BMC Evolutionary Biology

BioMed Central

Research article

Open Access

Birth and death of protein domains: A simple model of evolution explains power law behavior

Georgy P Karev[1], Yuri I Wolf[1], Andrey Y Rzhetsky[2], Faina S Berezovskaya[3] and Eugene V Koonin*[1]

Address: [1]National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD 20894, USA, [2]Columbia Genome Center, Columbia University, 1150 St. Nicholas Avenue, Unit 109, New York, NY 10032, USA and [3]Department of Mathematics, Howard University, 2400 Sixth Str., Washington D.C., 20059, USA

E-mail: Georgy P Karev - karev@ncbi.nlm.nih.gov; Yuri I Wolf - wolf@ncbi.nlm.nih.gov; Andrey Y Rzhetsky - ar345@columbia.edu; Faina S Berezovskaya - fberezovskaya@howard.edu; Eugene V Koonin* - koonin@ncbi.nlm.nih.gov

*Corresponding author

Published: 14 October 2002

BMC Evolutionary Biology 2002, **2**:18

Received: 3 September 2002
Accepted: 14 October 2002

This article is available from: http://www.biomedcentral.com/1471-2148/2/18

Abstract

Background: Power distributions appear in numerous biological, physical and other contexts, which appear to be fundamentally different. In biology, power laws have been claimed to describe the distributions of the connections of enzymes and metabolites in metabolic networks, the number of interactions partners of a given protein, the number of members in paralogous families, and other quantities. In network analysis, power laws imply evolution of the network with preferential attachment, i.e. a greater likelihood of nodes being added to pre-existing hubs. Exploration of different types of evolutionary models in an attempt to determine which of them lead to power law distributions has the potential of revealing non-trivial aspects of genome evolution.

Results: A simple model of evolution of the domain composition of proteomes was developed, with the following elementary processes: i) domain birth (duplication with divergence), ii) death (inactivation and/or deletion), and iii) innovation (emergence from non-coding or non-globular sequences or acquisition via horizontal gene transfer). This formalism can be described as a birth, death and innovation model (BDIM). The formulas for equilibrium frequencies of domain families of different size and the total number of families at equilibrium are derived for a general BDIM. All asymptotics of equilibrium frequencies of domain families possible for the given type of models are found and their appearance depending on model parameters is investigated. It is proved that the power law asymptotics appears if, and only if, the model is balanced, i.e. domain duplication and deletion rates are asymptotically equal up to the second order. It is further proved that any power asymptotic with the degree not equal to -1 can appear only if the hypothesis of independence of the duplication/deletion rates on the size of a domain family is rejected. Specific cases of BDIMs, namely simple, linear, polynomial and rational models, are considered in details and the distributions of the equilibrium frequencies of domain families of different size are determined for each case. We apply the BDIM formalism to the analysis of the domain family size distributions in prokaryotic and eukaryotic proteomes and show an excellent fit between these empirical data and a particular form of the model, the second-order balanced linear BDIM. Calculation of the

parameters of these models suggests surprisingly high innovation rates, comparable to the total domain birth (duplication) and elimination rates, particularly for prokaryotic genomes.

Conclusions: We show that a straightforward model of genome evolution, which does not explicitly include selection, is sufficient to explain the observed distributions of domain family sizes, in which power laws appear as asymptotic. However, for the model to be compatible with the data, there has to be a precise balance between domain birth, death and innovation rates, and this is likely to be maintained by selection. The developed approach is oriented at a mathematical description of evolution of domain composition of proteomes, but a simple reformulation could be applied to models of other evolving networks with preferential attachment.

Background

Sequencing of numerous genomes from all walks of life, including multiple representatives of diverse lineages of bacteria, archaea and eukaryotes, creates unprecedented opportunities for comparative-genomic studies [1–3]. One of the mainstream approaches of genomics is comparative analysis of protein or domain composition of predicted proteomes [2,4,5]. These studies often concentrate on domains rather than entire proteins because many proteins have variable multidomain architectures, particularly in complex eukaryotes (throughout this work, we use the term domain to designate a distinct evolutionary unit of proteins, which can occur either in the standalone form or as part of multidomain architectures; often but not necessarily, such a unit corresponds to a structural domain). As soon as genome sequences of bacteria became available, it has been shown that a substantial fraction of the genome of each species, from approximately one third in bacteria with very small genomes, to a significant majority in species with larger genomes, consists of families of paralogs, genes that evolved via gene duplication at different stages of evolution [6–9]. Again, a comprehensive analysis of paralogous relationships between genes is probably best performed at the level of individual protein domains, first, because many proteins share only a subset of common domains, and second, because domains can be conveniently and with a reasonable accuracy detected using available collections of domain-specific sequence profiles [10–12]. Comparisons of domain repertoires revealed both substantial similarities between different species, particularly with respect to the relative abundance of house-keeping domains, and major differences [4,5]. The most notable manifestation of such differences is lineage-specific expansion of protein/domain families, which probably points to unique adaptations [13,14]. Furthermore, it has been demonstrated that more complex organisms, e.g. vertebrates, have a greater variety of domains and, in general, more complex domain architectures of proteins than simpler life forms [1,2].

Lineage-specific expansions and gene loss events detected as the result of comparative analysis of the domain compositions of different proteomes have been examined mostly at a qualitative level, in terms of the underlying biological phenomena, such as adaptation associated with expansion or coordinated loss of functionally linked sets of genes [15]. A complementary approach involves quantitative comparative analysis of the frequency distributions of proteins or domains in different proteomes. Several studies pointed out that these distributions appeared to fit the power law: $P(i) \approx ci^{-\gamma}$ where $P(i)$ is the frequency of domain families including exactly i members, c is a normalization constant and γ is a parameter, which typically assumes values between 1 and 3 [16–19]. Obviously, in double-logarithmic coordinates, the plot of P as a function of i is a straight line with a negative slope. Power laws appear in numerous biological, physical and other contexts, which seem to be fundamentally different, e.g. distribution of the number of links between documents in the Internet, the population of towns or the number of species that become extinct within a year. The famous Pareto law in economics describing the distribution of people by their income and the Zipf law in linguistics describing the frequency distribution of words in texts belong in the same category [20–29]. Recent studies suggested that power laws apply to the distributions of a remarkably wide range of genome-associated quantities, including the number of transcripts per gene, the number of interactions per protein, the number of genes or pseudogenes in paralogous families and others [30].

Power law distributions are scale-free, i.e. the shape of the distribution remains the same regardless of scaling of the analyzed variable. In particular, scale-free behavior has been described for networks of diverse nature, e.g. the metabolic pathways of an organism or infectious contacts during an epidemic spread [20,25–27]. The principal pattern of network evolution that ensures the emergence of power distributions (and, accordingly, scale-free properties) is preferential attachment, whereby the probability of a node acquiring a new connection increases with the number of connections this node already has.

However, a recent thorough study suggested that many biological quantities claimed to follow power laws, in fact, are better described by the so-called generalized Pareto

function: $P(i) = c(i+a)^{-\gamma}$ where a is an additional parameter [31]. Obviously, although at $i >> a$, a generalized Pareto distribution becomes indistinguishable from a power law, at small i, it deviates significantly, the magnitude of the deviation depending on the value of a. Furthermore, unlike power law distributions, generalized Pareto distributions do not show scale-free properties.

The importance of the analysis of frequency distributions of domains or proteins lies in the fact that distinct forms of such distributions can be linked to specific models of evolution. Therefore, by exploring the distributions, inferences potentially can be made on the mode and parameters of genome evolution. For this purpose, the connections between domain frequency distributions and evolutionary models need to be explored theoretically within a maximally general class of models. In this work, we undertake such a mathematical analysis using simple models of evolution, which include duplication (birth), elimination (death) and *de novo* emergence (innovation) of domains as elementary processes (hereinafter BDIM, birth- death- innovation models). All asymptotics of equilibrium frequencies of domain families of different size possible for BDIM are identified and their dependence on the parameters of the model is investigated. In particular, analytical conditions on birth and death rates that produce power asymptotics are determined. We prove that the power law asymptotics appears if, and only if, the model is balanced, i.e. domain duplication and deletion rates are asymptotically equal up to the second order, and that any power asymptotic with the degree not equal to -1 can appear only if the assumption of independence of the duplication/deletion rates on the size of a domain family is rejected. We apply the developed formalism to the analysis of the frequency distributions of domains in individual prokaryotic and eukaryotic genomes and show a good fit of these data to a particular version of the model, the second-order balanced linear BDIM.

Results and Discussion

Mathematical theory and model

Fundamental definitions and assumptions

A genome is treated as a "bag" of coding sequence for protein domains, which we simply call ***domains*** for brevity. Domains are treated as independently evolving units disregarding the dependence between domains that tend to belong to the same multidomain protein. Each domain is considered to be a member of a ***family*** (including single-member families). We consider three types of elementary evolutionary events: i) domain ***birth,*** which generates a new member within a family; the principal mechanism of birth is duplication with divergence but additional mechanisms may be considered, including acquisition of a family member from a different species via horizontal gene transfer [32], ii) domain ***death,*** which results from domain inactivation and/or deletion, and c) domain ***innovation,*** which generates a new family with one member. Innovation may occur via horizontal gene transfer from another species, via domain evolution from a non-coding sequence or a sequence of a non-globular protein, or via major change of a domain from a pre-existing family after a duplication, which makes the relationship between the given domain and its family of origin undetectable (this latter process formally combines domain birth, death and innovation in a single event). The innovation rate (ν), is considered constant for a given genome. The rates of elementary events are considered to be independent of time (i. e. only homogeneous models are considered) and of the nature (structure, biological function etc.) of individual families.

In a finite genome, the maximal number of domains in a family cannot exceed the total number of domains and, in reality, is probably much smaller; let N be the maximal possible number of domain family members. We consider classes of domain families, which have only one common feature, namely the number of members (Fig. 1). Let f_i be the number of domain families in i-th class, i.e. families that are represented by exactly i domains in the given genome, $i = 1,2,...N$. Birth of a domain in a family of class i results in the relocation of this family from class i to class $i+1$ (decrease of f_i and increase of f_{i+1} by 1). Conversely, death of a domain in a family of class i relocates the family to class i-1; death of a domain in class 1 results in the elimination of the corresponding family from the given genome, this being the only considered mechanism of family death. We consider time to be continuous and suppose it very unlikely that more than one elementary event occur during a short time interval; formally, the probability that more than one event occurs during an interval Δt is $o(\Delta t)$.

The formulation of the model

The simple BDIM

Let us formulate the following **independence assumption**: i) all elementary events are independent of each other; ii) the rates of individual domain birth (λ) and death (δ) do not depend on i (number of domains in a family). Under this assumption, the instantaneous rate, at which a domain family leaves class i, is proportional to i and the following simple BDIM describes the evolution of such a system of domain family classes:

$$df_1(t)/dt = -(\lambda + \delta) f_1(t) + 2\delta f_2(t) + \nu$$

$$df_i(t)/dt = (i - 1)\lambda f_{i-1}(t) - i(\lambda + \delta) f_i(t) + (i + 1)\delta f_{i+1}(t) \text{ for } 1<i<N, \quad (2.1)$$

$$df_N(t)/dt = (N - 1)\lambda f_{N-1}(t) - N\delta f_N(t).$$

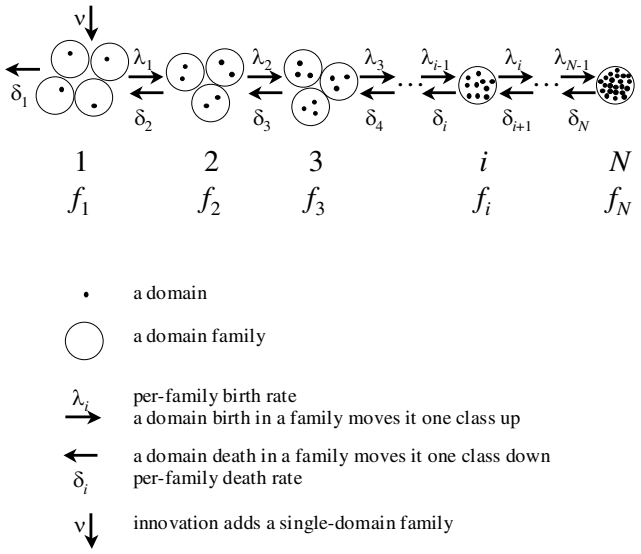

Figure 1
Domain dynamics and elementary evolutionary events under BDIM.

Similar models have been considered previously in several different contexts [33 v. 1, ch. 17, 34]. We will see in 3.2 that the solution of model (2.1) evolves to equilibrium, with a unique distribution of domain family sizes, $f_i \sim (\lambda/\delta)^i/i$; in particular, if $\lambda = \delta$, then $f_i \sim 1/i$. Thus, under the simple BDIM, if the birth rate equals the death rate, the abundance of a domain class is inversely proportional to the size of the families in this class. When the observations do not fit this particular asymptotic (as observed in several studies on distributions of protein family sizes), a different, more general model needs to be developed.

The Master BDIM

A more general BDIM emerges when the independence assumption is abandoned. Instead of constructing specific hypotheses regarding the dependence between the elementary events, let us simply suppose that the domain birth and death rates for a family of class i do not necessarily show proportionality to i. For the general case, we designate these rates, respectively, λ_i and δ_i; in the specific case of the simple BDIM (2.1), $\lambda_i = \lambda i$ and $\delta_i = \delta i$. Then we have the following ***master BDIM***:

$$df_1(t)/dt = -(\lambda_1 + \delta_1)f_1(t) + \delta_2 f_2(t) + \nu$$

$$df_i(t)/dt = \lambda_{i-1}f_{i-1}(t) - (\lambda_i + \delta_i)f_i(t) + \delta_{i+1}f_{i+1}(t) \text{ for } 1<i<N, \quad (2.2)$$

$$df_N(t)/dt = \lambda_{N-1}f_{N-1}(t) - \delta_N f_N(t).$$

Let $F(t) = \sum_{i=1}^{N} f_i(t)$ be the total number of domain families at instant t; it follows from (2.2) that

$$dF(t)/dt = \nu - \delta_1 f_1(t) \quad (2.3)$$

The system (2.2) has an equilibrium solution $f_1,...f_N$ defined by the equality $df_i(t)/dt = 0$ for all i; this solution is described below under Proposition 1. Accordingly, there exists an equilibrium solution of equation (2.3), which we will designate F_{eq} (the total number of domain families at equilibrium). At equilibrium, $\nu = \delta_1 f_1$, i.e. the processes of innovation and death of *single* domains (more precisely, the death of domain families of class 1, i.e. singletons) are balanced.

We can rewrite the model (2.2) in terms of the frequency of a domain family of class i $p_i(t) = f_i(t)/F(t)$. Let $x(t) = \gamma(t)/Y(t)$; then

$$dx/dt = [d\gamma/dt\ /\gamma - dY/dt\ /Y]\ x.$$

Applying this identity to $p_i(t)$ and rewriting equation (2.3) in the form

$$[dF(t)/dt]/F(t) = \nu/F(t) - \delta_1 p_1(t) \quad (2.3')$$

we obtain the following model for frequencies of the domain family (*master BDIM for frequencies*), which is equivalent to (2.2):

$$dp_1(t)/dt = -(\lambda_1 + \delta_1)p_1(t) + \delta_2 p_2(t) + \nu/F(t) - (\nu/F(t) - \delta_1 p_1(t))p_1(t), \quad (2.4)$$

$$dp_i(t)/dt = \lambda_{i-1}p_{i-1}(t) - (\lambda_i + \delta_i)p_i(t) + \delta_{i+1}p_{i+1}(t) - (\nu/F(t) - \delta_1 p_1(t))\ p_i(t) \text{ for } 1<i<N,$$

$$dp_N(t)/dt = \lambda_{N-1}p_{N-1}(t) - \delta_N p_N(t) - (\nu/F(t) - \delta_1 p_1(t))]\ p_N(t).$$

System (2.4) should be solved together with equation (2.3).

The Master BDIM and Markov processes

Let us note that system (2.4) for frequencies is *non-linear*, so it is not a system of Kolmogorov equations for state probabilities of any homogeneous Markov process. Let us further suppose that a genome had ample time to arrive at an equilibrium with respect to the total number of domain families, such that $F(t) = F_{eq}$. This does not imply $dp_i(t)/dt = 0$ or $df_i(t)/dt = 0$; in other words, the system might rearrange the frequencies of individual families, although the total number of families remains stable. If $F(t) = F_{eq}$, the master system (2.4) turns into

$$d\,p_1(t)/dt = -(\lambda_1 + \delta_1)\,p_1(t) + \delta_2 p_2(t) + \nu/F_{eq} \qquad (2.5)$$

$d\,p_i(t)/dt = \lambda_{i-1}p_{i-1}(t) - (\lambda_i + \delta_i)\,p_i(t) + \delta_{i+1}p_{i+1}(t)$ for $1<i<N$,

$d\,p_N(t)/dt = \lambda_{N-1}p_{N-1}(t) - \delta_N p_N(t)$.

System (2.5) can be rewritten as a matrix equation

$d\boldsymbol{p}(t)/dt = \boldsymbol{p}(t)Q$,

where $\boldsymbol{p}(t) = \{p_1(t),...p_N(t)\}$ and the matrix $Q = (q_{ij})$ is defined by equalities

$q_{11} = -(\lambda_1 + \delta_1) + \nu/F_{eq}$, $q_{21} = \delta_2 + \nu/F_{eq}$, $q_{s1} = \nu/F_{eq}$ for all $s > 2$;

$q_{i-1,i} = \lambda_{i-1}$, $q_{i,i} = -(\lambda_i + \delta_i)$, $q_{i+1,i} = \delta_{i+1}$, $q_{k,i} = 0$ for all k, $|i-k| > 1$, $i = 2,...N$-1,

$q_{N-1,N} = \lambda_{N-1}$, $q_{N,N} = -\delta_N$.

It is easy to see that the sum of elements of each row (except for the first one) of the matrix Q is equal to $\nu/F_{eq} > 0$. Therefore the matrix Q cannot be a matrix of transition rates for any Markov process (the sum of elements of each row of a matrix of transition rates for Markov process with continuous time should be non-positive [33 v. 1, ch.17, s. 8, 35 v. 2, ch. 3, s. 2]; in other words, there is no Markov process with continuous time and state space $\{1,2,...N\}$ whose state probabilities satisfy system (2.5).

Thus, neither the initial BDIMs (2.1) or (2.2) nor the equilibrium model (2.5) can be described by any Markov process with continuous time.

Remark. If, in system (2.5), $\nu = 0$, then this system turns into a system of state probabilities for a Markov birth and death process with continuous time.

Equilibrium in BDIMs

Equilibrium sizes and frequencies of the domain family system

Let us suppose that the genome had ample time to arrive at a complete equilibrium state, in which not only $dF(t)/dt = 0$, but also $df_i(t)/dt = 0$ for all i. Thus, the equilibrium sizes of domain families f_i satisfy the system

$$-(\lambda_1 + \delta_1)\,f_1 + \delta_2 f_2 + \nu = 0,$$

$$\lambda_{i-1}f_{i-1} - (\lambda_i + \delta_i)f_i + \delta_{i+1}f_{i+1} = 0 \text{ for } 1<i<N, \qquad (3.1)$$

$$\lambda_{N-1}f_{N-1} - \delta_N f_N = 0.$$

It should be emphasized that the master model does not assign *a priori* the value of F_{eq}; this value has to be computed depending on the model parameters.

The following statement is central for further analysis.

Proposition 1. *The master BDIM (2.2) has a unique equilibrium state* $(f_1,...f_N)$*, which is the sole solution* of *system (3.1):*

$$f_1 = \nu/\delta_1$$

$$f_i = \nu \prod_{j=1}^{i-1} \lambda_j \Big/ \prod_{j=1}^{i} \delta_j \text{ for all } i = 2,...N. \qquad (3.2)$$

The unique equilibrium state (3.2) is globally asymptotically stable.

In addition (*formally assuming* $\prod_{j=1}^{i-1} \lambda_j = 1$ for $i = 1$),

$$\sum_{i=1}^{N} F_{eq} = \nu \Big(\prod_{j=1}^{i-1} \lambda_j \Big/ \prod_{j=1}^{i} \delta_j \qquad (3.3)$$

This proposition ascertains that all evolutionary trajectories of the system (2.2) exponentially (with respect to time) approach the equilibrium state (3.2). The proof is given in the Mathematical Appendix.

Remark. Let us denote the ratio of the birth rate to the innovation rate

$$G(N) \equiv \sum_{i=1}^{N-1} \lambda_i f_i/\nu,$$

and the ratio of the death rate to the innovation rate

$$I(N) \equiv \sum_{i=1}^{N} \delta_i f_i/\nu.$$

Then, according to Proposition 1, for any BDIM in equilibrium,

$$G(N) - I(N) = \sum_{i=1}^{N-1} \prod_{j=1}^{i} \lambda_j/\delta_j - \sum_{i=1}^{N-1} \prod_{j=1}^{i} \lambda_j/\delta_j - 1 = -1.$$

The principal goal of the treatment that follows is the analysis of the asymptotic behavior of equilibrium frequencies and sizes of domain families $(f_1,...f_N)$ at large N. We will differentiate two cases of asymptotic behavior according to the following

Definition. Let $\{q_i\}$, $\{s_i\}$ be sequences of real numbers; let us denote $q_i \cong s_i$ if $\lim q_i/s_i = 1$ and $q_i \sim s_i$ if $\lim q_i/s_i = c = const$ and $0<c<\infty$. We will also use this notation for finite but sufficiently long sequences.

Equilibrium frequencies for the simple BDIM

Let us apply Proposition 1 to the simple BDIM (2.1) with $\lambda_i = \lambda i$, $\delta_i = \delta i$.

Definition. A simple BDIM is *balanced* if $\theta = \lambda/\delta = 1$, i.e. if the rates of individual domain birth and death are equal.

Let us recall that a random discrete variable ξ has the *logarithmic* distribution with parameter $\theta < 1$ if

$$P(\xi = i) = \theta^i/i\ [-\ln(1-\theta)]^{-1},\ i = 1,2,...$$

A random variable ξ has the *truncated logarithmic* distribution with parameter θ if

$$P(\xi = i) = C_n\,\theta^i / i,\ i = 1,2,...n,\ C_n = 1/\sum_{i=1}^{N} \theta^j/j.$$

Then, we have

Proposition 2.

1) *For any simple BDIM (2.1)*

$$f_i = (\nu/\delta)\theta^{i-1}/i = (\nu/\lambda)\theta^i/i, \qquad (3.4)$$

$$F_{eq} = \sum_{i=1}^{N} f_i = \nu/\delta \sum_{j=1}^{N} \theta^{j-1}/j, \qquad (3.5)$$

and

$$p_i = (1/F_{eq})(\nu/\delta)\theta^{i-1}/i = (\theta^i/i) / \sum_{j=1}^{N} \theta^j/j \qquad (3.6)$$

that is, the equilibrium frequencies have the truncated logarithmic distribution if $\theta < 1$.

2) *If a simple BDIM is balanced, then*

$$F_{eq} = \nu/\delta \sum_{j=1}^{N} 1/j, \qquad (3.7)$$

and for all $i = 1,2,...N$

$$p_i = \nu/\delta F_{eq}/i = \Big(\sum_{j=1}^{N} 1/j\Big)^{-1} / i. \qquad (3.8)$$

The proof is given in the Mathematical Appendix.

Thus, a simple BDIM can have equilibrium frequencies only of the form $p_i = C\theta^i/i$, where $C = const$ and θ is the distribution parameter. In particular, the equilibrium frequencies for a balanced simple BDIM have the power distribution with the degree equal to -1.

Simple methods exist for preliminary graphical estimation of the single distribution parameter θ [36 ch. 7, s. 7]. We will prove in the following section that, if we observe a power asymptotic for empirically observed equilibrium frequencies, then (assuming that the system can be described by a BDIM), the rates λ_i and δ_i should be asymptotically equal at large i. If, additionally, the degree of the asymptotic is *not equal* to -1, then the system dynamics *cannot* be described by a *simple* BDIM. In this case, it is necessary to consider more general models, such as the Master BDIM (2.2).

Asymptotic behavior of equilibrium frequencies of a Master BDIM: Main Theorems

Let us consider the master BDIM (2.2); we showed in 3.1 that its equilibrium frequencies are the solution of the system

$$-(\lambda_1 + \delta_1)p_1 + \delta_2 p_2 + \nu/F_{eq} = 0, \qquad (3.9)$$

$$\lambda_{i-1}p_{i-1} - (\lambda_i + \delta_i)p_i + \delta_{i+1}p_{i+1} = 0 \text{ for } 1<i<N,$$

$$\lambda_{N-1}p_{N-1} - \delta_N p_N = 0.$$

The following theorem gives all possible types of asymptotic behavior of the equilibrium frequencies and defines the connections between these asymptotics depending on model parameters. In particular, if there is no information on the exact form of dependence of the rates of birth and death of domains on the size of a domain family, the theorem can be used to qualitatively describe the dynamics of the asymptotic behavior of the equilibrium frequencies.

We will prove that the asymptotic behavior of a solution of system (3.9) is completely defined by the asymptotic relation between λ_i and δ_i. More precisely, let us define a function $\chi(i) = \lambda_{i-1}/\delta_i$; we consider only functions of power growth, i.e. $\chi(i) \sim i^s$ at $i \to \infty$ for a real s. We will see that this is not a serious restriction because the most realistic situations correspond to the case of $s = 0$. So, let us suppose that, for large i, the following expansion is valid:

$$\chi(i) \equiv \lambda_{i-1}/\delta_i = i^s\,\theta\,(1+a/i + O(1/i^2)) \qquad (3.10)$$

where s, a are real numbers and $\theta > 0$. Evidently, if $s \neq 0$, $\chi(i)$ tends either to 0 ($s < 0$) or to ∞ ($s > 0$) with the increase of i.

Definition. Let us refer to a BDIM (2.2), (3.10) as

i. non-balanced, if $s \neq 0$;

ii. first-order balanced, if $s = 0$ and $\theta \neq 1$, i.e.

$$\lambda_{i-1}/\delta_i = \theta\,(1+a/i + O(1/i^2)) \text{ at large } i; \qquad (3.11)$$

iii. second-order balanced, if $s = 0$, $\theta = 1$ and $a \neq 0$, i.e.

$$\lambda_{i-1}/\delta_i = 1 + a/i + O(1/i^2)) \text{ for large } i; \qquad (3.12)$$

iv. high-order balanced, if $s = 0$, $\theta = 1$ and $a = 0$, i.e.

$\lambda_{i-1}/\delta_i = 1 + O(1/i^2))$ for large i.

We will show that the first three coefficients, s, θ and a, of asymptotic expansion (3.10) for $\chi(i) = \lambda_{i-1}/\delta_I$ exactly specify all possible asymptotic behaviors of BDIM equilibrium frequencies.

Theorem 1. *The equilibrium frequencies p_i of BDIM (2.2) have the following asymptotics*

i. if the model is non-balanced, then

$p_i \sim \Gamma(i)^s \theta^i i^a$, where $\Gamma(i)$ is the Γ-function;

ii. if the model is first-order balanced, then

$p_i \sim \theta^i i^a$;

iii. if the model is second-order balanced, then

$p_i \sim i^a$;

iv. if the model is high-order balanced, then

$p_i \sim 1$

The proof is given in the Mathematical Appendix. The classification of BDIM according to the order of balance is illustrated in Fig. 2 and the asymptotics for different types of BDIMs are shown in Fig. 3.

It follows from this theorem that, if a BDIM is non-balanced, then its equilibrium frequencies p_i (and equilibrium family sizes f_i) increase or decrease extremely fast (hyper-exponentially) with the increase of i. In contrast, if a BDIM has a non-zero order of balance, asymptotic behavior is observed.

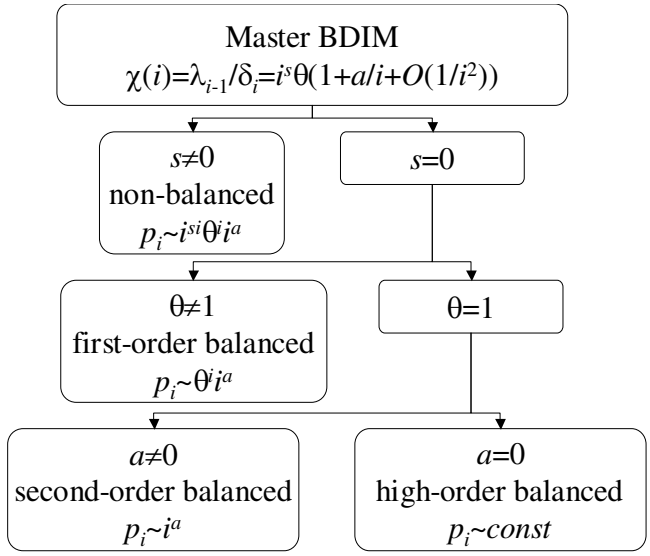

Figure 2
Different orders of balance in BDIMs.

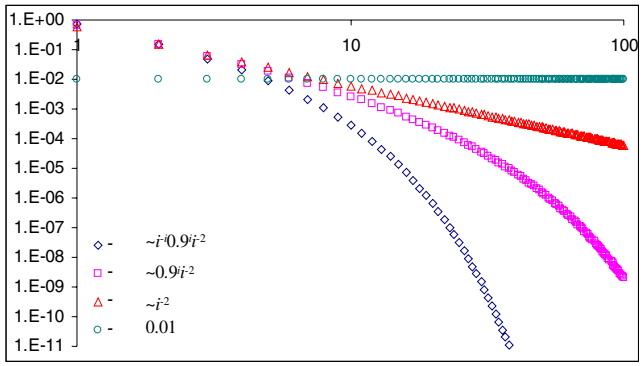

Figure 3
Asymptotics of equilibrium distributions for balanced BDIMs of different orders.

Let us recall that a random discrete variable ξ has the *Pascal* (or *negative binomial*) *distribution* with parameters (r,q), $r > 0$, $0 < q < 1$, if $P(\xi = k) = \Gamma(r+k)/[\Gamma(r)\,\Gamma(1+k)]\,(1-q)^r q^k$[36]. We will say that sequence $\{p_i\}$ follows (or asymptotically has) a discrete probabilistic distribution $\{q_i\}$ if $p_i \sim q_i$ for large enough i.

Corollary 1.*For a first-order balanced BDIM with* $\theta < 1$,

i. if $a > -1$, *the equilibrium frequencies* p_i *follow Pascal distribution with parameters* $(a+1,\theta)$;

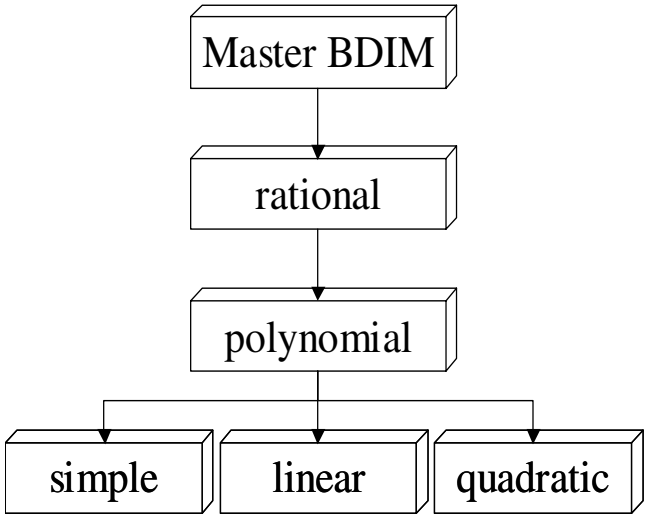

Figure 4
The hierarchy of BDIM types.

ii. if a = -1, the equilibrium frequencies follow truncated logarithmic distribution with parameter θ;

iii. if a = 0, the equilibrium frequencies follow geometric distribution with parameter θ.

The following implication of Theorem 1 is of principal interest.

Corollary 2. *Equilibrium frequencies of a BDIM have a power asymptotic behavior if and only if the BDIM is second-order balanced.*

Corollary 3. *For high-order balanced BDIM, if* $\lambda_{i-1}/\delta_i = 1$ *for all i, the only possible distribution of equilibrium frequencies is uniform,* $p_i = const$ *for all i. Moreover, even if* $\lambda_{i-1}/\delta_i = 1 + O(1/i^2)$, *the equilibrium frequencies asymptotically tend to the uniform distribution.*

Rational BDIM

Rational models comprise a general class of BDIM (Fig. 4), for which the asymptotic behavior of the equilibrium frequencies and equilibrium sizes of domain families can be completely investigated.

Let us suppose that the birth and death rates are of the form

$$\lambda_i = \lambda\, P(i) = \lambda \prod_{k=1}^{n} (i + a_k)^{\wedge}\alpha_k, \qquad (4.1)$$

$$\delta_i = \delta\, Q(i) = \delta \prod_{k=1}^{m} (i + b_k)^{\wedge}\beta_k$$

for $i > 0$, where λ, δ are positive constants, α_k, β_k are real and a_k, b_k are non-negative for all $k = 1,...N$.

We will refer to BDIM (2.2.), (4.1) as *rational* BDIM.

It is known that a wide class of mathematical functions can be well approximated by rational functions of the form (4.1) (see, e.g. [37]).

Specific cases of the rational BDIM are *simple* BDIM with $P(i) = i$, $Q(i) = i$, *linear* BDIM with $P(i) = i + a_1$, $Q(i) = i + b_1$, where a_1, b_1 are constants, and polynomial BDIM, if $P(i)$ and $Q(i)$ are polynomials on i.

The following theorem describes all possible asymptotic behaviors of the equilibrium frequencies of a rational BDIM. Let us denote

$$\theta = \lambda/\delta,$$

$$\eta = \sum_{k=1}^{n} \alpha_k - \sum_{k=1}^{m} \beta_k,$$

$$\rho = \sum_{k=1}^{n} a_k\alpha_k - \sum_{k=1}^{m} b_k\beta_k,$$

$$\beta = \sum_{k=1}^{m} \beta_k.$$

Theorem 2. *The equilibrium sizes of domain families of a rational BDIM have the following asymptotics*

$$f_i \cong C\nu/\lambda\, \Gamma(i)^{\eta}\, \theta^i i^{\rho-\beta} \qquad (4.2)$$

where the constant

$$C = \prod_{k=1}^{m} (\Gamma(1 + b_k)^{\wedge}\beta_k / \prod_{k=1}^{n} \Gamma(1 + a_k)^{\wedge}\alpha_k. \qquad (4.3)$$

The proof is given in the Mathematical Appendix.

Corollary 1. *If* η = 0, *then the rational BDIM is first-order balanced and the sequence of equilibrium numbers of domain families* $\{f_i\}$ *has a power-exponential asymptotics*

$$f_i \cong C\nu/\lambda\, \theta^i i^{\rho-\beta}. \qquad (4.4)$$

In particular, if $\rho - \beta > -1$, the equilibrium frequencies p_i follow the Pascal distribution with parameters $(\rho - \beta + 1, \theta)$;

if $\rho - \beta = -1$, *then frequencies p_i follow the truncated logarithmic distribution;*

if $\rho - \beta = 0$, *then frequencies p_i follow the geometric distribution.*

Corollary 2. *The equilibrium sizes of domain families f_i and equilibrium frequencies p_i for a rational BMID have the power asymptotics if and only if $\eta = 0$ and $\lambda = \delta$, i.e. the BDIM is second-order balanced, in which case*

$$f_i \cong C\nu/\lambda\, i^{\rho-\beta}. \qquad (4.5)$$

Formula (4.4) gives the asymptotics for the equilibrium sizes of domain families f_i and, accordingly, for the total number of families F_{eq}. The exact expressions for these quantities are given in the proofs of Theorem 2 and Lemma (see Mathematical Appendix).

Proposition 3.

i. The equilibrium sizes of domain families f_i of a balanced (first or higher order) rational BDIM are

$$f_i = C\nu/\delta\theta^{i-1} \prod_{k=1}^{n} [(\Gamma(i + a_k))^{\alpha_k}] / \prod_{k=1}^{m} [(\Gamma(i + 1 + b_k))^{\beta_k}]$$ for all $i = 1,2,...$

where

$$C = \prod_{k=1}^{m} [(\Gamma(1 + b_k))^{\beta_k}] / \prod_{k=1}^{n} (\Gamma(1 + a_k))^{\alpha_k}].$$

ii. The total number of domain families at equilibrium is

$$F_{eq} = C\nu/\delta\Big(\sum_{j=1}^{N} \theta^{j-1} \prod_{k=1}^{n} (\Gamma(j + a_k))^{\alpha_k} / \prod_{k=1}^{m} (\Gamma(j + 1 + b_k))^{\beta_k}\Big).$$

For the rational, second-order balanced BDIM, the ratio of the birth rate to the innovation rate is

$$G(N) = \sum_{i=1}^{N-1} \theta^i \prod_{k=1}^{m} [\Gamma(i + 1 + a_k)/\Gamma(1 + a_k)]^{\alpha_k} / [\Gamma(i + 1 + b_k) / \Gamma(1 + b_k)]^{\beta_k}.$$

The asymptotic formulas for equilibrium frequencies of rational BDIM could be considered as particular cases of the corresponding formulas of general theorem 1. Proposition 3 allows one to calculate the constants in the corresponding asymptotic formulas for the sizes of domain families for a rational BDIM. If only equilibrium frequencies are analyzed, the values of these constants become irrelevant because they contract. However, if the actual values of f_i and F_{eq} are of interest, the values of the constants are required.

Properties of the main types of rational BDIM

Simple BDIM

As shown above, a simple BDIM can have equilibrium frequencies only of the form $p_i = C\theta^i/i$, $C = const$;in particular, if the distribution parameter $\theta < 1$, we get the (truncated) logarithmic distribution. Logarithmic distributions are seen in many biological contexts, e.g., the distribution of species by the number of individuals in populations or, what is more relevant, the distribution of protein folds by the number of families per fold [38]. Thus, a simple BDIM could be potentially used for modeling the dynamics of biological systems with a logarithmic distribution of equilibrium densities. We examine this possibility in greater detail starting with the case $\lambda = \delta$ (second-order balanced simple BDIM).

We can extract from Proposition 2 some additional information, which could be helpful for estimating the model parameters. It is known that

$\sum_{i=1}^{N} 1/i = \ln N + C_E + O(1/N)$, where C_E is the *Euler* constant, $C_E = 0.5772157...$

More precisely, the approximation

$\sum_{i=1}^{N} 1/i = \ln N + C_E + N^{-1}/2 - N^{-2}/12$ has an error less then 10^{-6} for $N > 10$. Thus, from (3.7), we obtain an interesting formula

$$F_{eq} \cong (\nu/\delta)\,[\ln N + C_E] \qquad (5.1)$$

This means that, in the equilibrium state of the system, the total number of domain families grows only slowly ($\sim \ln N$) with the increase of the maximal number (N) of domains in a family (which is equal to the maximal possible number of domain family size classes).

Furthermore, according to equation (2.3), in the equilibrium state of a simple BDIM $\nu/\delta = f_1$, so we have

$F_{eq} / f_1 \cong \ln N + C_E$ (5.2)

Formula (5.1) can be used for estimating the model parameters on the basis of empirical data.

In the more general case $\lambda \neq \delta$, we can also obtain an estimate of the rate of innovation ν. If $\lambda < \delta$ ($\theta < 1$), then the series in the right part of (3.5) quickly converges,

$$\sum_{i=1}^{N} \theta^{i-1}/i \to -\ln(1-\theta)/\theta,$$

so $-\ln(1-\theta)/\theta$ is a good approximation for the sum $\sum_{i=1}^{N} \theta^{i-1}/i$ for large N. Then

$$F_{eq} = (\nu/\delta) \sum_{i=1}^{N} \theta^{i-1}/i = (\nu/\lambda) \sum_{i=1}^{N} \theta^{i}/i \cong \nu/\lambda\ (-\ln(1-\theta)),$$

and

$$\nu/\delta = F_{eq}\, \theta/(-\ln(1-\theta)). \quad (5.3)$$

Taking into account that $\nu/\delta = f_1$ (2.3), we have a relation

$$F_{eq}/f_1 \cong -\ln(1-\theta)/\theta, \quad (5.4)$$

which allows the parameter θ to be estimated on the basis of empirical data.

If N can be estimated independently and is not very large, we can use more exact relations:

$$\sum_{i=1}^{N} \theta^{i}/i \cong -\ln(1-\theta) + Ei(-N(1-\theta)) - N^{-1}/2 + N^{-2}/12.$$

where the function $Ei(u) = \int_{-\infty}^{u} e^{x}/x\,dx$.

Further, if $(1-\theta)N$ is small (i.e., θ is very close to 1), then the approximation

$$\sum_{i=1}^{N} \theta^{i}/i \cong C_E - N(1-\theta)$$

has an error less then $[N(1-\theta)]^2/4$ and, in this case,

$$F_{eq}/f_1 \cong (C_E - N(1-\theta))/\theta. \quad (5.5)$$

If $(1-\theta)N$ is large, then the following inequalities provide simple bounds for $F_{eq}/f_1 = \sum_{i=1}^{N} \theta^{i-1}/i$:

$$-(\ln(1-\theta)/\theta - \theta^N/[(N+1)(1-\theta)] < \sum_{i=1}^{N} \theta^{i-1}/i < -\ln(1-\theta)/\theta - \theta^N[1/(N+1) - \theta/(N+2)]. \quad (5.6)$$

For the simple BDIM, the ratio of the rate of duplications to the innovation rate is

$$G(N) = \sum_{i=1}^{N-1} \lambda_i f_i/\nu = \sum_{i=1}^{N-1} \theta^{i} = \theta(1-\theta^{N-1})/(1-\theta),$$

so $G(N) \to \infty$ if $\theta > 1$ and $G(N) \to 1/(1-\theta)$ if $\theta < 1$ at $N \to \infty$.

If the simple BDIM is the 2nd order balanced, $\theta = 1$, then $G(N) = N - 1$.

Thus, for the simple, second-order balanced BDIM, the number of duplications per time unit is N-1 times greater than the number of innovations.

The total number of domains in the equilibrium state for the simple BDIM is

$$M(N) = \sum_{i=1}^{N} if_i = \nu/\lambda\theta(1-\theta^N)/(1-\theta).$$

If a simple BDIM is second-order balanced, then $G(N) = \nu/\lambda\, N$.

Linear BDIM

We saw that the assumption of independence of birth and death rates of individual domains on each other and on the size of domain families is incompatible with any power distribution of the equilibrium frequencies with the degree not equal to -1. The simplest case of a BDIM, which can have, depending on the parameters, three types of asymptotic behavior described by Theorem 1 (excluding the first one, hyper-exponential, which corresponds to a non-balanced BDIM; all linear BDIMs are balanced) and, in particular, any power asymptotics, is a model with linear birth and death rates of the form:

$\lambda_i = \lambda\ (i + a)$, $\delta_i = \delta\ (i + b)$, where a and b are constants. (5.7)

The parameters a and b account, in the simplest possible form, for the deviation of the domain birth and death

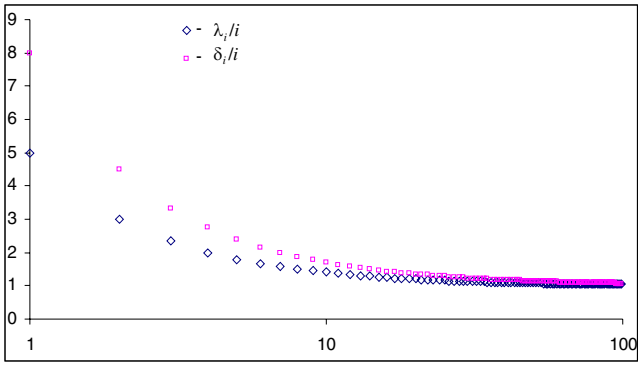

Figure 5
Dependence of per domain birth and death rates on the domain family size for the second-order balanced linear BDIM.

rates from those under the independence assumption. More precisely, according to (5.7), the average birth rate per domain in a family of size i is $\lambda_i/i = \lambda + \lambda a/i$. So, for small i, the average birth rate is close to $\lambda + \lambda a$, whereas, for large i, it tends to λ. Similarly, the average death rate changes from $\delta + \delta b$ in a small family to the limit value δ in a large family. Thus, if a and b are positive (which seems to be the case for the available data; see below), both the birth rate and the death rate per domain decrease with the increase of the class number (size of the respective domain families); conversely, if a and b are negative, these rates increase with the class number (Fig. 5).

Corollary 1 of Theorem 2 implies that equilibrium frequencies p_i of a linear BDIM have asymptotics

$$p_i \sim \theta^i i^{a-b-1}, \text{ where } \theta = \lambda/\delta. \qquad (5.8)$$

In particular, if $\lambda \neq \delta$ and $a = b$, the linear BDIM is first-order balanced and the equilibrium frequencies p_i follow the logarithmic distribution (in this case, the linear BDIM is asymptotically equivalent to the simple BDIM). If $\lambda = \delta$, the linear BDIM is second-order balanced and the equilibrium frequencies p_i follow the power distribution

$$p_i \sim i^{a-b-1}. \qquad (5.9)$$

Thus, the dependence of the domain frequency on the family size is actually determined by the difference $a - b$. If $a > b$, the birth rate decreases faster than the death rate with the increase of family size, i. e. there seems to be a "competition" between domains in a family; in contrast, if $a < b$, the death rate drops faster, i.e. a "synergy" between domains appears to exist (Fig. 4).

More detailed information can be obtained using Proposition 4:

i) for a first-order balanced linear BDIM, the equilibrium sizes f_i of domain families are

$$f_i = c\nu/\delta\theta^{i-1}\Gamma(i + a)/(\Gamma(i + 1 + b)) \text{ for all } i$$

where

$$c = \Gamma(1 + b)/\Gamma(1 + a)$$

and the total number of domain families at equilibrium is

$$F_{eq} = c\nu/\delta[\sum_{j=1}^{N} \theta^{j-1}\Gamma(j + a) / (\Gamma(j + 1 + b)]. \qquad (5.10)$$

ii) for a second-order balanced linear BDIM ($\theta = 1$),

$$f_i = c_1\nu/\delta\,\Gamma(i + a)/\Gamma(i + 1 + b)$$

and

$$F_{eq} = c\nu/\delta(\sum_{j=1}^{N}\Gamma(j+a)/\big(\Gamma(j+1+b)\big) =$$
$$\nu/\delta(1-\frac{\Gamma(1+b)\Gamma(1+a+N)}{\Gamma(1+a)\Gamma(1+b+N)})/(b-a) \qquad (5.11)$$

According to (2.3), in the equilibrium state of a linear BDIM, $f_1 = \nu/\delta_1 = \nu/(\delta(1 + b))$ and so, for a *second-order balanced* linear BDIM, we have the formula

$$F_{eq}/f_1 = (1+b)(1-\frac{\Gamma(1+b)\Gamma(1+a+N)}{\Gamma(1+a)\Gamma(1+b+N)})/(b-a)$$

Suppose that equilibrium frequencies obtained from empirical data follow the power distribution $p_i \sim i^{-\gamma}$; in this case, $-\gamma$ is the slope of the empirical curve ($\ln f_i$ versus $\ln i$) and can be estimated from the data. Assuming that the system is well described by a linear BDIM, it follows from (5.9) that $a - b = 1 - \gamma$ and $\lambda = \delta$. Thus,

$$f_i = c\nu/\delta\,\Gamma(i + a)/\Gamma(i + a + \gamma), \text{ where } c = \Gamma(\gamma + a)/\Gamma(1 + a), \qquad (5.12)$$

$$F_{eq} = c\nu/\delta\sum_{j=1}^{N}\Gamma(j+a)/\Gamma(j+a+\gamma) =$$
$$\nu/\delta(1-\frac{\Gamma(1+a+N)\Gamma(a+\gamma)}{\Gamma(a+\gamma+N)\Gamma(1+a)})/(\gamma-1)$$

and

$$F_{eq}/f_1 = (a+k)(1-\frac{\Gamma(1+a+N)\Gamma(a+\gamma)}{\Gamma(a+\gamma+N)\Gamma(1+a)})/(\gamma-1)$$

where a is the single free parameter.

For the linear second-order balanced BDIM, the ratio of the birth rate to the innovation rate is

$$G(N) = \sum_{i=1}^{N-1}\lambda_i f_i/\nu = \sum_{i=1}^{N-1}\Gamma(1+b)/\Gamma(1+a)(i+a)\Gamma(i+a)/(\Gamma(i+1+b) =$$
$$\sum_{i=1}^{N-1}\Gamma(1+b)/\Gamma(1+a)\Gamma(i+1+a)/(\Gamma(i+1+b) =$$
$$\frac{\Gamma(1+a+N)\Gamma(1+b)}{(1+a-b)\Gamma(b+N)\Gamma(1+a)} + \frac{1+a}{b-a-1}$$

if $1 + a - b \neq 0$. As

$$\frac{\Gamma(1+a+N)\Gamma(1+b)}{(1+a-b)\Gamma(b+N)\Gamma(1+a)} \cong \frac{\Gamma(1+b)}{(1+a-b)\Gamma(1+a)}N^{1+a-b},$$
$$G(N) \to \frac{1+a}{b-a-1}$$

if $1 + a - b < 0$ and $G(N)\to\infty$ if $1 + a - b > 0$ at $N\to\infty$.

The case $1 + a - b = 0$ (slope of the asymptote in double logarithmic coordinates equal to $a - b - 1 = -2$) is a critical one.

In this case,

$$G(N) = \sum_{i=1}^{N-1} \Gamma(1+b)/\Gamma(b)\,\Gamma(i+b)/(\Gamma(i+1+b) =$$

$$\sum_{i=1}^{N-1} b_1/(i+b) = b\,[\text{PolyGamma}(0, b+N) - \text{PolyGamma}(0, b+1)].$$

Accordingly, $G(N)\to\infty$ at $N\to\infty$.

The total number of domains in the equilibrium state for a second-order balanced linear BDIM is

$$M(N) = \sum_{i=1}^{N} i f_i =$$
$$\nu/\delta\sum_{i=1}^{N} i\Gamma(1+b)/\Gamma(1+a)\Gamma(i+a)/(\Gamma(i+1+b) =$$
$$\nu/\delta[\frac{a}{a-b} + \frac{1+a}{b-a-1} + \frac{\Gamma(1+b)}{\Gamma(1+b+N)}(\frac{\Gamma(2+N+a)}{(1+a-b)\Gamma(1-a)} - \frac{\Gamma(1+N+a)}{(a-b)\Gamma(a)})].$$

If the slope of the asymptote $\gamma = -1$, the linear second-ordered BDIM shows the same asymptotic behavior as a simple BDIM (2.1), but behaves differently at small i. If $\gamma \neq -1$, the system cannot be described by a *simple* BDIM even asymptotically, but can be described by a linear BDIM. As indicated above, in this case, the average per-domain birth and death rates depend on the size of the domain family and the difference a-b characterizes this dependence.

Quadratic BDIM

The linear BDIM takes into account the dependence of average birth and death rates of individual domains on the size of domain family, but does not imply a specific form of interaction between domains. Let us consider the simplest, pairwise interaction, which leads to $\lambda_i \sim i^2$ and/or $\delta_i \sim i^2$, i.e. one or both rates are polynomials on i of the second degree. If these degrees are different (i.e., $\lambda_i \sim i$ and $\delta_i \sim i^2$), then the corresponding BDIM is non-balanced and equilibrium frequencies have hyper-exponential asymptotics. Thus, let

$$\lambda_i = \lambda\,(i^2 + r_1 i + r_2),\ \delta_i = \delta\,(i^2 + q_1 i + q_2), \qquad (5.13)$$

where r_k, q_k, $k = 1,2$ are constants (such that λ_i, δ_i are positive for all i) or

$$\lambda_i = \lambda\,(i + a_1)(i + a_2),$$

$$\delta_i = \delta\,(i + b_1)(i + b_2)$$

Then, $r_1 = a_1 + a_2$, $q_1 = b_1 + b_2$, and

$$\chi(i) = \lambda_{i-1}/\delta_i = \theta\,(1 + (r_1 - q_1 - 2)/i + O(1/i^2)),$$

where $\theta = \lambda/\delta$.

According to theorem 3 and Proposition 3, the quadratic BDIM with rates (5.13) has equilibrium sizes of domain families

$f_i = c_2\ \nu/\delta\ \theta^{i-1}\ \Gamma\ (i + a_1)\ \Gamma\ (i + a_2)\ /\ (\Gamma\ (i + 1 + b_1)\ (\Gamma\ (i + 1 + b_2)) \cong c_2\nu/\delta\ \theta^{i-1} i^{\rho-2}$ (5.14)

where $\rho = r_1 - q_1$ and the constant $c_2 = [(\Gamma\ (1+b_1)\ \Gamma\ (1+b_2)]\ /\ [\Gamma\ (1+a_1)\ \Gamma\ (1+a_2)]$, and the total number of domain families at equilibrium

$$F_{eq} = c_2\nu/\delta\ \left(\sum_{j=1}^{N} \theta^{j-1}\ \Gamma(j+a_1)\ \Gamma(j+a_2)\ /\ (\Gamma(j+1+b_1)\ (\Gamma(j+1+b_2))\right). \quad (5.15)$$

Note that the asymptotic behavior of frequencies p_i does not depend on free coefficients r_2, q_2 in (5.13), but only on θ and r_1-q_1 (as follows from (5.14)), although the values of f_i are proportional to the constant c_2, which could depend on the free coefficients r_2, q_2. Let us consider the case $r_2 = q_2 = 0$ in more detail.

If only square terms are present in the expressions for the birth and death rates, $\lambda_i = \lambda i^2$, $\delta_i = \delta i^2$, then $a_k = b_k = 0$, k = 1,2 and so $c_2 = 1$, $f_i = \nu/\delta\ \theta^{i-1}/i^2$ and $F_{eq} = \nu/\delta\ \sum_{j=1}^{N} \theta^{j-1}/j^2$. So at $N\to\infty$

$$F_{eq} \cong\ \nu/\delta\ \sum_{j}\ \theta^{j-1}\ /\ j^2 = \nu/\lambda\ \text{Polylog}(2,\theta) \quad (5.16)$$

where Polylog is a special function, $\text{Polylog}(k,x) = \sum_{j=1}^{\infty} x^j/j^k$.

According to (3.2), $f_1 = \nu/\delta_1$; for this particular case of quadratic BDIM, $f_1 = \nu/\delta$ and

$$F_{eq}/f_1 \cong\ \text{Polylog}(2,\theta). \quad (5.17)$$

Formula (5.17) allows estimating parameter θ from empirical data if N is large enough.

More precisely, $F_{eq} = \nu/\lambda\ \sum_{j=1}^{N} \theta^j/j^2 = \nu/\lambda\ (\text{Polylog}(2,\theta) - \theta^{1+N}\ \text{LerchPhi}(\theta,2,1+N))$, where LerchPhi is a special function (these and other special functions used below can be computed using program packages Mathematika or Maple).

If, additionally, θ = 1 (the BDIM is second-order balanced), then

$$f_i = (\nu/\delta)/i^2 = f_1/i^2 \quad (5.18)$$

and, at large N

$$F_{eq} \cong\ \nu/\delta\ \pi^2/6\ \cong\ 1.645\ \nu/\delta = 1.645 f_1. \quad (5.19)$$

From formulas (5.8), (5.15), we can extract some additional information, which could be helpful for estimating the model parameters at relatively small N. Let us recall definitions of some special functions.

The digamma function φ(z) is logarithmic derivative of Γ(z), φ(z) = Γ'(z)/Γ(z).

The function PolyGamma(n,z) is n^{th} derivative of φ(z), PolyGamma(n,z) = $d^n\phi(z)/dz^n$, such that φ(z)= PolyGamma(0,z).

It is known that

$$\sum_{i=1}^{N}\ 1/i^2 = \pi^2/6\text{-PolyGamma}(1,1+N),$$

Thus we have

$$F_{eq} = \nu/\delta\ \sum_{j=1}^{N}\ 1/j^2 = \nu/\delta\ [\pi^2/6\text{-PolyGamma}(1,1+N)] \quad (5.20)$$

$$F_{eq}/f_1 = \pi^2/6\text{-PolyGamma}(1,1+N),$$

which can be used for estimating unknown parameters of the model.

The values of PolyGamma(1,x) are tabulated and can be computed using standard program packages; for a rough preliminary estimate, PolyGamma(1,x) = $1/x+1/2x^2+O(1/x^4)$.

If linear terms are also present in the quadratic BDIM, $\lambda_i = \lambda\ (i^2+a_1 i)$, $\delta_i = \delta\ (i^2+b_1 i)$, then

$$f_i = c_2\nu/\delta\ \theta^{i-1}/i\ \Gamma\ (i+a_1)/\Gamma\ (i+1+b_1)$$

where $c_2 = \Gamma\ (1+b_1)/\Gamma\ (1+a_1)$; $F_{eq} = \Sigma f_i$ can be computed using special functions. In particular, if the BDIM is second-order balanced, θ = 1, then

$f_i = c_2\nu/\delta\ \Gamma\ (i+a_1)\ /\ (i\ \Gamma\ (i+1+b_1))$.

For this variant of the model, $f_1 = \nu/\delta_1 = \nu/(\delta(1+b_1))$, and so

$$f_i = f_1 \frac{\Gamma(2+b_1)\Gamma(i+a_1)}{i\Gamma(1+a_1)\Gamma(i+1+b_1)}.$$

Polynomial BDIMs

The quadratic models take into account the dependence of birth and death rates of individual domains on the simplest, pairwise interactions. If interactions of higher orders are postulated, $\lambda_i \sim P_n(i)$ and/or $\delta_i \sim Q_m(i)$, where $P_n(i)$, $Q_m(i)$ are polynomials on i of the n-th and m-th degrees. Again, if the degrees n and m are different, then the BDIM is non-balanced and equilibrium frequencies have hyper-exponential asymptotics. Thus, let $n = m$,

$$\lambda_i = \lambda R\ (i) = \lambda \sum_{k=0}^{m} r_k i^{m-k},\ \delta_i = \delta Q(i) = \delta \sum_{k=0}^{m} q_k i^{m-k} \quad (5.21)$$

where r_k, q_k are constants and $r_0 = q_0 = 1$. We suppose, of course, that $R(i)$, $Q(i)$ are positive for all integer i. Note that, in this case, $\chi\ (i) \equiv \lambda_{i-1}/\delta_i = \theta\ (1+(r_1 - q_1 - m)/i+O(1/i^2))$, where $\theta = \lambda/\delta$. We will suppose that $\theta \le 1$.

According to Theorem 3, the polynomial BDIM with rates (5.21) has equilibrium sizes of domain families with *power-exponential asymptotics*

$$f_i \sim \theta^i i^{\rho-m} \quad (5.22)$$

where $\rho = r_1 - q_1$.

In particular, if $\rho - m > -1$, the equilibrium frequencies p_i follow the Pascal distribution with parameters $(\rho - m + 1, \theta)$;

if $\rho - m = -1$, the equilibrium frequencies p_i follow the (truncated) logarithmic distribution;

if $\rho - m = 0$, the equilibrium frequencies p_i follow the geometric distribution;

if $\lambda = \delta$, the polynomial BDIM is second-order balanced and the equilibrium frequencies p_i follow the power distribution

$$p_i \sim i^{\rho-m}. \quad (5.23)$$

Note that the degree of the power distribution (5.23) depends only on m, the common degree of the polynomials (5.21), and on ρ, the difference between the coefficients r_1 and q_1, and does not depend on other coefficients. In particular, if $r_1 = q_1$, then $p_i \sim i^{-m}$. This relation could be interpreted as follows: if the first two coefficients of polynomial rates λ_i and δ_i are equal, then the degree of the power distribution (5.19) is equal to the "order of interactions" of domains.

Formula (5.22) can be refined. Let $R(i) = \prod_{k=1}^{m} (i+a_k)$, $Q(i) = \prod_{k=1}^{m} (i+b_k)$.

Then (see Proposition 3) the equilibrium numbers of domain families f_i of the polynomial BDIM (5.18) are

$$f_i = C\nu/\delta\theta^{i-1} \prod_{k=1}^{m} [\Gamma(i+a_k)/\Gamma(i+1+b_k)]$$

where $C = \prod_{k=1}^{m} [\Gamma(1+b_k)/\Gamma(1+a_k)]$, and the equilibrium total number of domain families

$$F_{eq} = C\nu/\delta \sum_{j=1}^{N} \theta^{j-1} \prod_{k=1}^{m} [\Gamma(j+a_k)/\Gamma(j+1+b_k)].$$

For the polynomial model $f_1 = \nu/\delta_1 = \nu/(\delta \sum_{k=0}^{m} q_k)$, so

$$F_{eq}/f_1 = C \sum_{j=1}^{N} \theta^{j-1} \prod_{k=1}^{m} (\Gamma(j+a_k)/\Gamma(j+1+b_k))/ \sum_{k=0}^{m} q_k.$$

This formula can be used for estimating the model parameters.

For the polynomial second-order balanced BDIM, the ratio of the death rate to the innovation rate is

$$G(N) = \sum_{i=1}^{N-1} \lambda_i f_i/\nu = (\prod_{k=1}^{m} \Gamma(1+b_k)/\Gamma(1+a_k))$$

$$\sum_{i=1}^{N-1} \prod_{k=1}^{m} \Gamma(i+1+a_k)/\Gamma(i+1+b_k) =$$

$$\sum_{i=1}^{N-1} \prod_{k=1}^{m} [\Gamma(i+1+a_k)/\Gamma(1+a_k)]/[\Gamma(i+1+b_k)/\Gamma(1+b_k)].$$

Table 1: Domain families in sequenced genomes and parameters of the best-fit second-order balanced linear BDIM

	No. of ORFs in genome	No. of detected domain families	No. of detected domains	No. of ORFs with RPS-BLAST hits	Maximum size of a family	f_1 (f'_1)	*a*	*b*	*k*	$\nu/\delta = \nu/\lambda$	$G = \sum_{i=1}^{N-1} \lambda_i f_i/\nu$
Sce[b]	6340	1080	4575	3331	130	420 (436)	1.55	3.27	-2.72	1861.8	[3.28..3.53]
Dme	13605	1405	11734	7262	335	426 (435)	1.62	2.79	-2.17	1648.2	[8.44..15.50]
Cel	20524	1418	17054	11090	662	423 (421)	1.13	2.03	-1.89	1273.0	[16.18..∞]
Ath	25854	1405	21238	15006	1535	270 (277)	3.80	4.98	-2.18	1657.7	[17.09..26.89]
Hsa	39883	1681	27844	16755	1151	298 (288)	5.16	6.43	-2.27	2136.2	[17.14..22.88]
Tma	1846	772	1683	1268	97	501 (499)	0.14	2.22	-3.08	1606.4	[1.04..1.06]
Mth	1869	693	1480	1150	43	438 (436)	0.12	2.00	-2.88	1305.3	[1.18..1.27]
Sso	2977	695	1950	1614	81	386 (385)	0.36	2.04	-2.68	1167.8	[1.83..2.00]
Bsu	4100	1002	3413	2502	124	507 (510)	0.48	2.01	-2.53	1534.6	[2.46..2.79]
Eco	4289	1078	3624	2765	140	523 (519)	0.84	2.54	-2.70	1837.0	[2.45..2.61]

a: f_1(f'_1), the observed (predicted) number of domains in class 1 (represented only once in the genome); *a*, *b*, parameters of the second-order balanced linear BDIM; *k*, slope of the power asymptotic; $\nu/\delta = \nu/\lambda$, ratio of the innovation rate to the per domain death (birth) rate; *G*, ratio of the innovation rate to the total (per genome) birth rate. b: Species abbreviations: Sce, *Saccharomyces cerevisiae*, Dme, *Drosophila melanogaster*, Cel, *Caenorhabditis elegans*, Ath, *Arabidopsis thaliana*, Hsa, *Homo sapiens*, Tma, *Thermotoga maritima*, Mth, *Methanothermobacter thermoautotrophicum*, Sso, *Sulfolobus solfataricus*, Bsu, *Bacillus subtilis*, Eco, *Escherichia coli*.

Approximation of the observed domain family size distributions in prokaryotic and eukaryotic genomes with different BDIMs

Having developed the mathematical theory of BDIMs, we sought to determine which of these models, if any, adequately described the empirical data on domain family size distribution. To identify the domain sets of domains encoded in each of the genomes, the CDD library of position-specific scoring matrices (PSSMs), which includes the domains from the Pfam and SMART databases, was used in RPS-BLAST searches [12] against the protein sequences from a set of completely sequenced eukaryotic and prokaryotic genomes [http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome]. The CDD domain library is partially redundant, so when the results obtained from individual PSSMs showed significant overlap (more than 50% of hits overlapped for more than 50% of their length), the corresponding domains were examined case-by-case for redundancy. PSSMs representing structurally similar domains and producing overlapping lists of hits were joined into "synonymy clusters".

The results of RPS-BLAST searches against the sets of protein sequences from individual genomes were interpreted as follows: non-overlapping hits to the same protein were treated independently; among overlapping hits, only the strongest one (lowest *E*-value) was recorded; all hits from a synonymy cluster were assigned to one representative domain family. The number of hits that a domain family had in a genome, with the cut-off of $E = 0.001$, was recorded as the number of domains of the given family encoded in the given genome. The CDD domain library certainly does not include all existing domains. In practice, domains from this collection were detected in >50% in each of the analyzed species, with the only exception of human, for which the analyzed protein set is likely to contain a substantial fraction of false predictions (Table 1).

To enable statistical analysis using the χ^2-method for the entire range of the data, including the sparsely distributed classes corresponding to large families, the data needed to be combined. For each genome, the observed domain family frequencies were grouped into bins, each containing at least 10 families; typically, bins corresponding to families with small number of members included a single size class (e.g. all single-member families, two-member families etc), whereas bins corresponding to large families may span hundreds of size classes, most of them empty. Theoretical distribution values for a bin combining observed data from *m*-th to *n*-th class were computed as $\sum_{i=m}^{n} f'_i$, where f'_i is the predicted number of families in the *i*-th class and depends on the model parameters. Since the model displays only a weak dependence on the maximum number of domains in a family (*N*), instead of including *N* as a model parameter, the sum $\sum_{i=1}^{i\max} f'_i$ (where i_{max} is the number of domains in the most abundant of the detected families), was normalized to equal the total number of families detected in the given genome (a re-

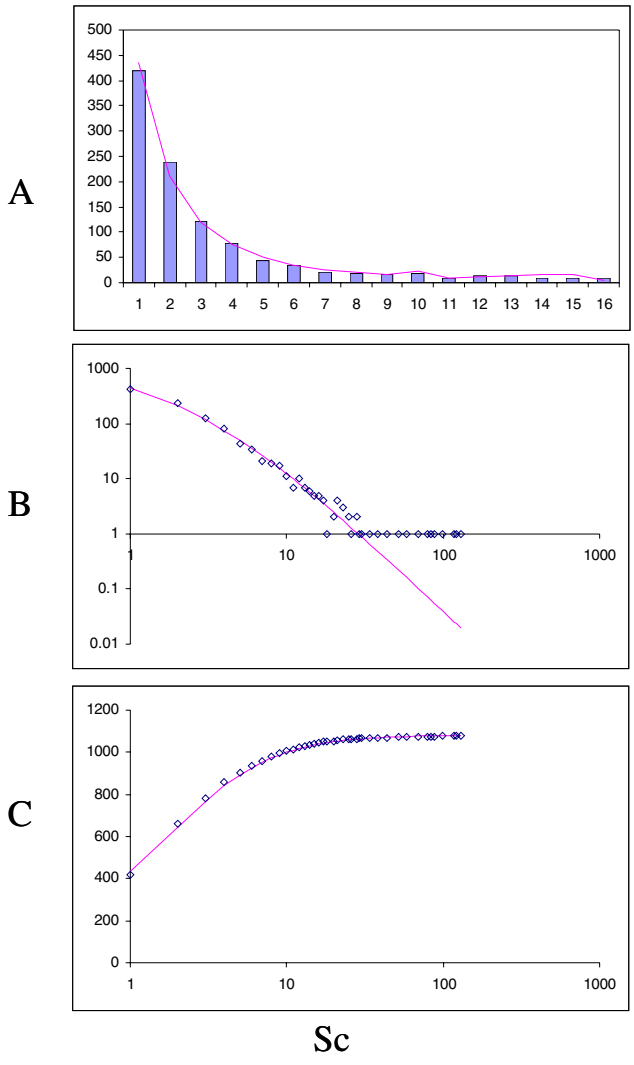

Figure 6
Fit of empirical domain family size distributions to the second-order balanced linear BDIM: the yeast *Saccharomyces cerevisiae*. A. Distribution of the size of domain families grouped into bins B. Domain family size distribution in double logarithmic coordinates. Magenta line: f_i = 11521Γ(*i*+1.55)/Γ(*i*+4.27) C. Cumulative distribution function of domain family size. The line shows the prediction of the second-order balanced linear BDIM.

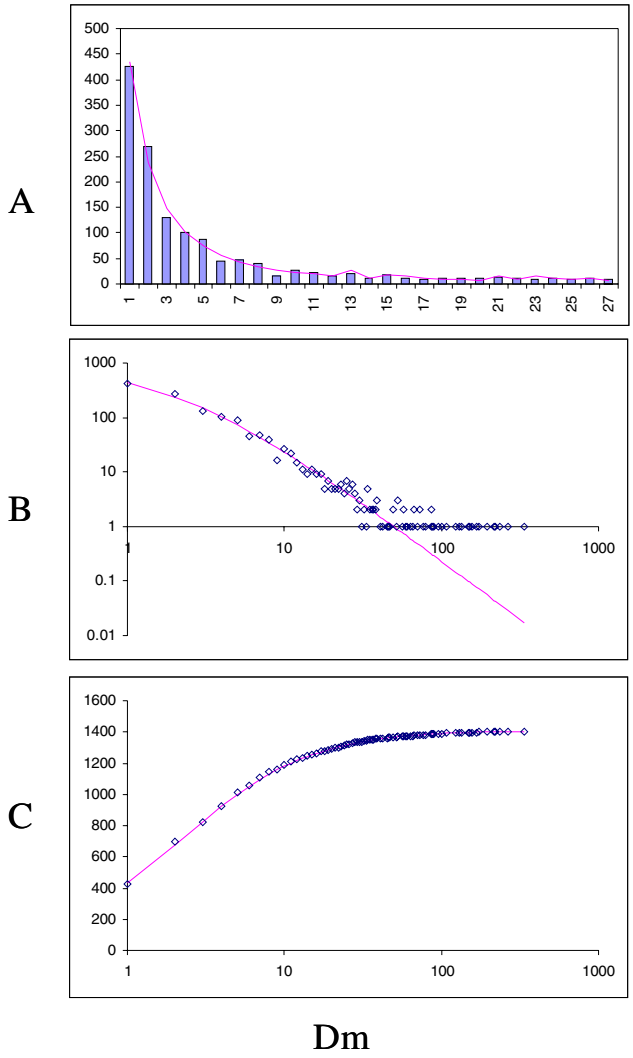

Figure 7
Fit of empirical domain family size distributions to the second-order balanced linear BDIM: the fruit fly *Drosophila melanogaster*. The panels and the designations are as in Fig. 6. B. Magenta line: f_i = 5258Γ(*i*+1.62)/Γ(*i*+3.79)

quirement for the χ^2 analysis). χ^2 values were computed to measure the quality of fit between the observed and the theoretical distributions. The distribution parameters (θ for the simple BDIM, *a* and *b* for the second-order balanced linear BDIM) were adjusted to minimize the χ^2 value.

The simplest model that resulted in a good fit to the observed domain family size distributions was the second-order balanced linear BDIM (Fig. 6,7,8,9,10,11,12,13,14,15). For all analyzed genomes, $P(\chi^2)$ for this model was >0.05, i.e. no significant difference between the model predictions and the observed data was detected. Considering the first-order balanced linear BDIM, which involves varying the parameter θ, did not result in a significant improvement of fit for any of the analyzed genomes (data not shown). In contrast, a fit to a truncated logarithmic distribution (prediction of a simple BDIM) failed for all genomes ($P(\chi^2) < 10^{-5}$; Fig. 16, 17, and data not shown). An exact power-law distribution, which is often used to approximate protein family frequency distributions, similarly failed to adequately fit the observed data, even when the most deviant class 1 fami-

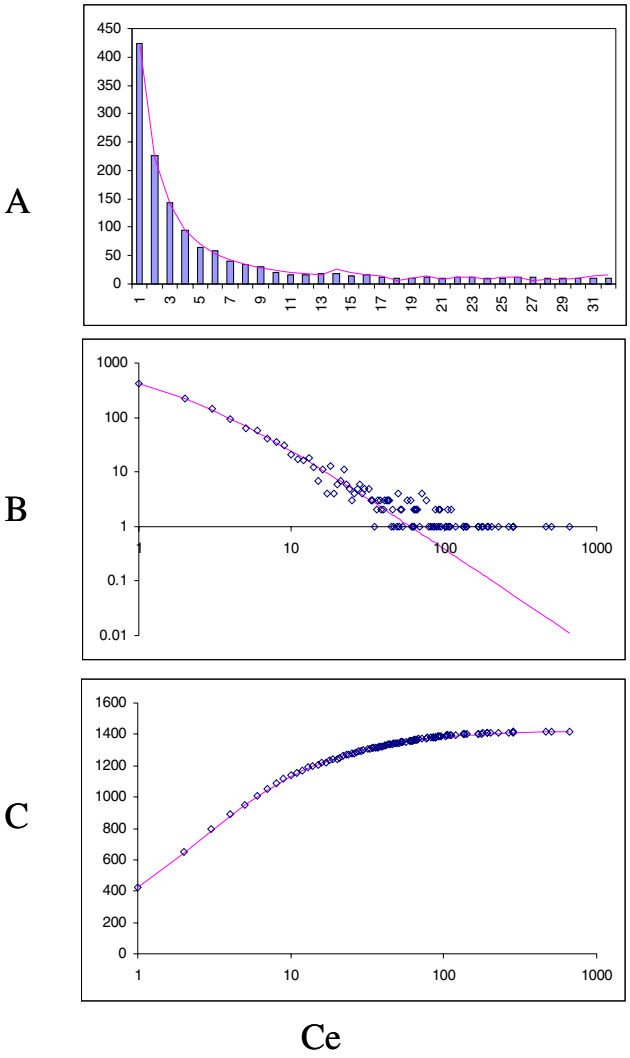

Figure 8
Fit of empirical domain family size distributions to the second-order balanced linear BDIM: the nematode worm *Caenorhabditis elegans*. The panels and the designations are as in Fig. 6. B. Magenta line: $f_i = 2453\Gamma(i+1.13)/\Gamma(i+3.03)$

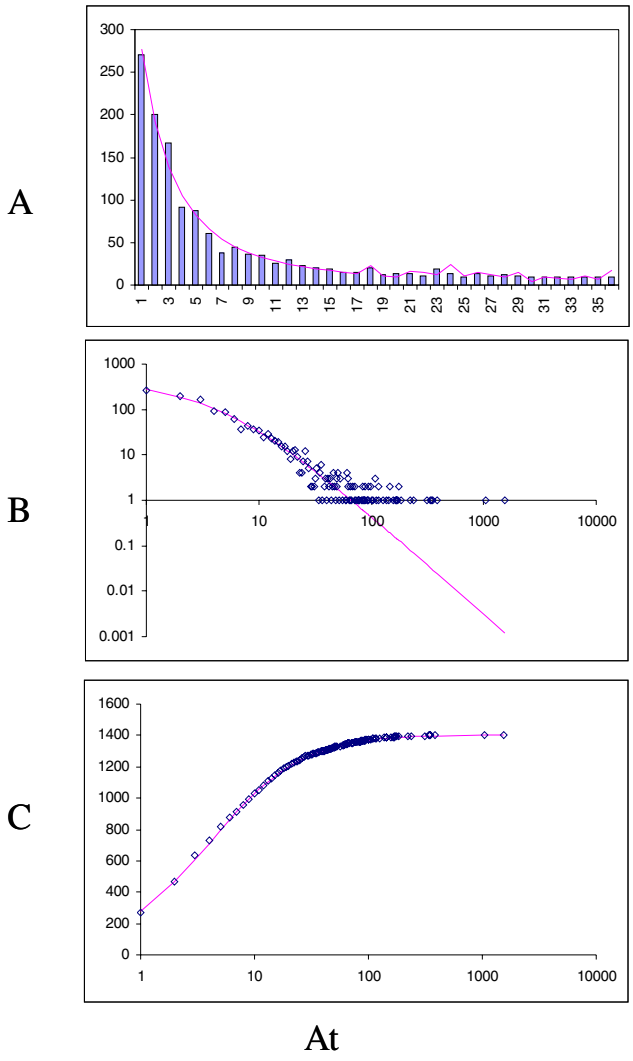

Figure 9
Fit of empirical domain family size distributions to the second-order balanced linear BDIM: the thale cress *Arabidopsis thaliana*. The panels and the designations are as in Fig. 6. B. Magenta line: $f_i = 10750\Gamma(i+3.80)/\Gamma(i+5.98)$

lies were excluded ($P(\chi^2) = 0.0013$ for *T. maritima*; $P(\chi^2) < 10^{-5}$ for the rest of the genomes; Fig. 16, 17 and data not shown). Notably, second-order balanced linear BDIM results in a correct prediction of the number of very large families, whereas simple BDIM systematically underestimates the number of families in the highest bins. Conversely, the power-law fit underestimates the slope of the best-fit line (in double logarithmic coordinates) compared to the asymptote of the linear BDIM prediction and, accordingly, significantly overestimates the number of families in the highest bins (Fig. 16, 17). These results are compatible with the recent observation that the domain family size distributions are better described by the generalized Pareto distribution than by power laws [31].

Fitting the observed domain family size distribution with the second-order balanced linear BDIM resulted in positive values of the parameters a and b, with $a < b$, for all analyzed genomes (Table 1). Accordingly, domain family size distributions in all cases asymptotically tend to the power law with the power $k < -1$ and, for all species with the exception of *C. elegans*, $k < -2$ (Table 1 and Fig. 8). As discussed above, this seems to indicate the existence of "synergy" between domains in a family whereby the likelihood of survival is greater for a domain that belongs to

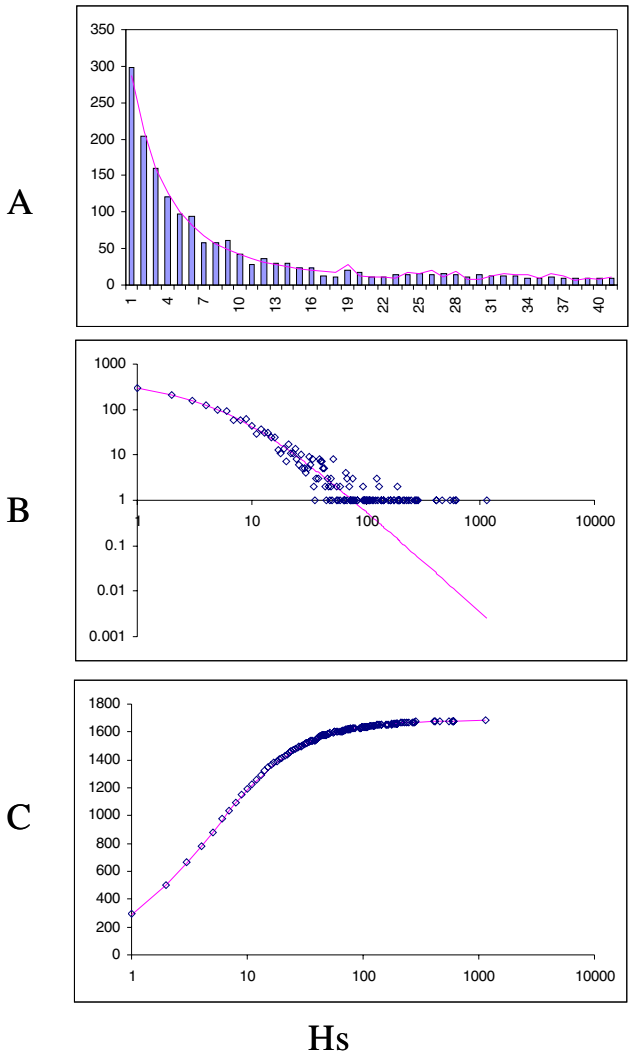

Figure 10
Fit of empirical domain family size distributions to the second-order balanced linear BDIM: *Homo sapiens*. The panels and the designations are as in Fig. 6. B. Magenta line: f_i = 22030Γ(*i*+5.16)/Γ(*i*+7.43)

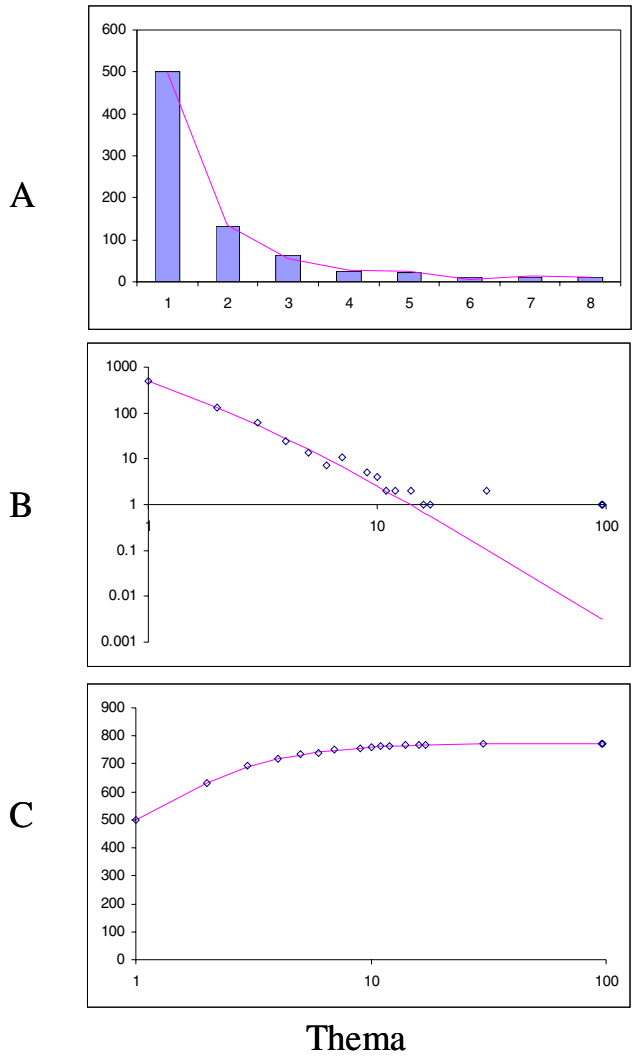

Figure 11
Fit of empirical domain family size distributions to the second-order balanced linear BDIM: the hyperthermophilic bacterium *Thermotoga maritima*. The panels and the designations are as in Fig. 6. B. Magenta line: f_i = 4256Γ(*i*+0.14)/Γ(*i*+3.22)

a large family than for a domain from a small family (Fig. 5). For all species, we find that the innovation rate is approximately three orders of magnitude greater than the per domain birth (death) rate. Accordingly, the total per genome birth (duplication) rate is comparable to but, typically, several times greater than the innovation rate (Table 1). The ratio of the per genome birth rate to the innovation rate increases with the number of genes in a genome or the number of detected domains, with nearly identical rates seen for small prokaryotic genomes and values as high as 20 for the largest plant and animal genomes (Table 1).

The data used to fit the BDIM typically included 50–60% of the proteins encoded in a given genome (Table 1); the remaining proteins were not represented by sufficiently similar domains in the current CDD collection. It cannot be ruled out that the fit would be significantly affected as a result of including all proteins encoded in the genome, in case the proteins currently not recognized in CDD searches have a family size distribution substantially different from that of the recognized ones. However, second-order balanced linear BDIM can accommodate considerable perturbations of the distribution through adjustment of the parameters, so we believe that this model is likely to approximate well also the size distribution of domain

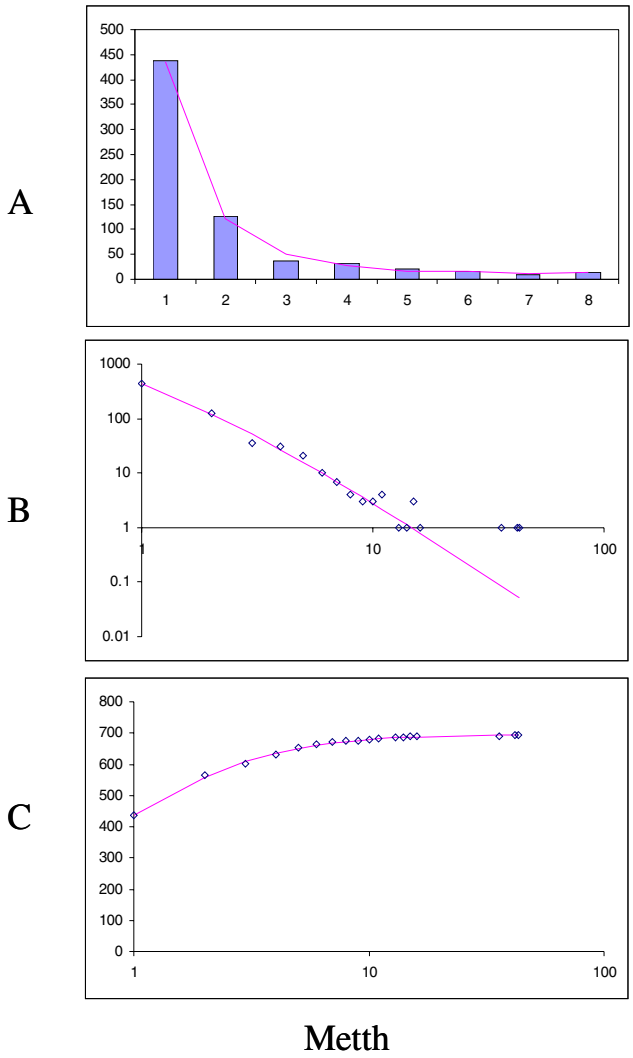

Figure 12
Fit of empirical domain family size distributions to the second-order balanced linear BDIM: the thermophilic euryarchaeon *Methanothermobacter thermautotrophicus*. The panels and the designations are as in Fig. 6. B. Magenta line: f_i = 2753Γ(*i*+0.12)/Γ(*i*+3.00)

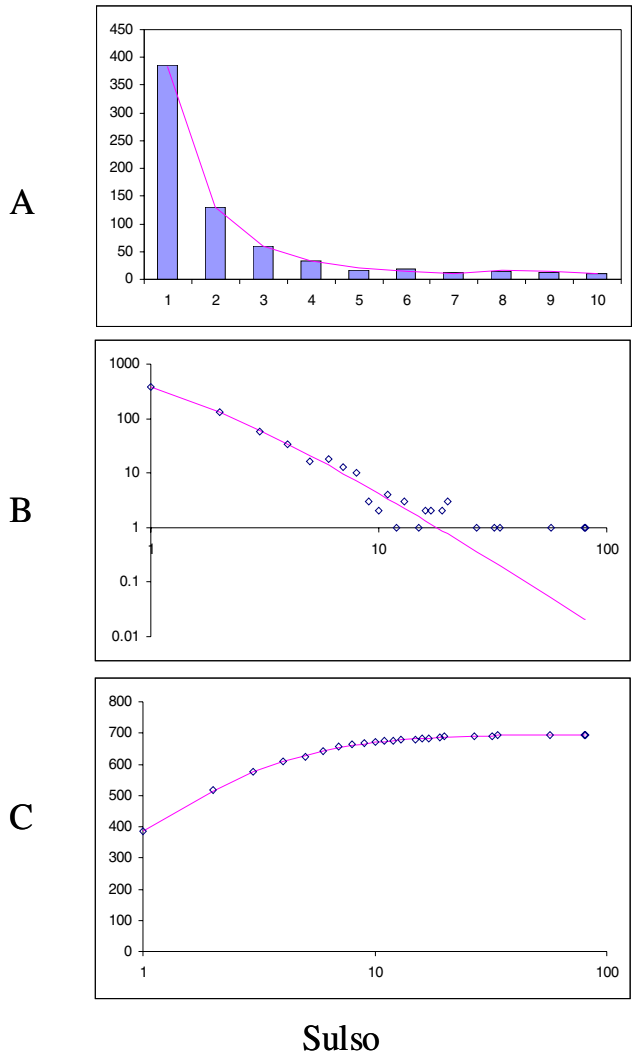

Figure 13
Fit of empirical domain family size distributions to the second-order balanced linear BDIM: the hyperthermophilic crenarchaeon *Sulfolobus solfataricus*. The panels and the designations are as in Fig. 6. B. Magenta line: f_i = 2714Γ(*i*+0.36)/Γ(*i*+3.04)

families for complete sets of proteins encoded in a genome. An alternative approach that at least partially circumvents the sampling problem involves analysis of all families of paralogs detectable using clustering by sequence similarity, with employing a predefined library of domains; this analysis is beyond the scope of the present work but may be a subject of further investigation.

General discussion and conclusions

Here, we presented a complete mathematical description of the size distribution of protein domain families encoded in genomes for simple but not unrealistic models of evolution, which include three types of events: domain duplication (birth), domain elimination (death), and domain innovation. In biological terms, innovation could involve gene acquisition via horizontal gene transfer, emergence of a new domain from a non-coding sequence or a non-globular protein sequence, or major modification of a domain obliterating its connection with a pre-existing family. Innovation via horizontal gene transfer appears to be particularly common in prokaryotes [32,39], which might account for the apparent higher relative innovation rate in prokaryotic genomes observed in the present analysis (Table 1).

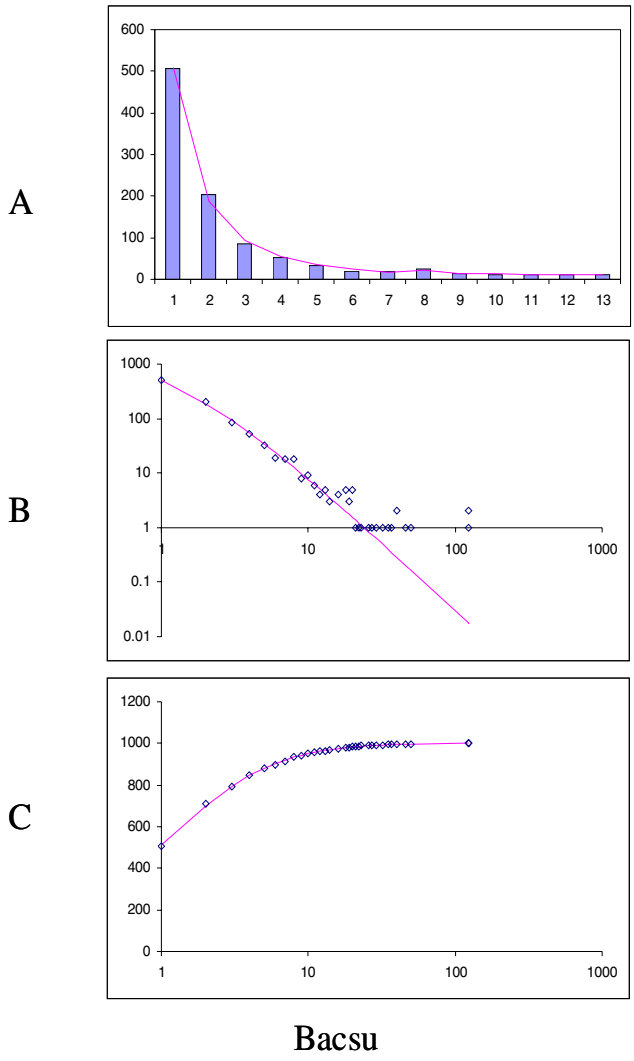

Figure 14
Fit of empirical domain family size distributions to the second-order balanced linear BDIM: the bacterium *Bacillus subtilis*. The panels and the designations are as in Fig. 6. B. Magenta line: $f_i = 3489\Gamma(i+0.48)/\Gamma(i+3.01)$

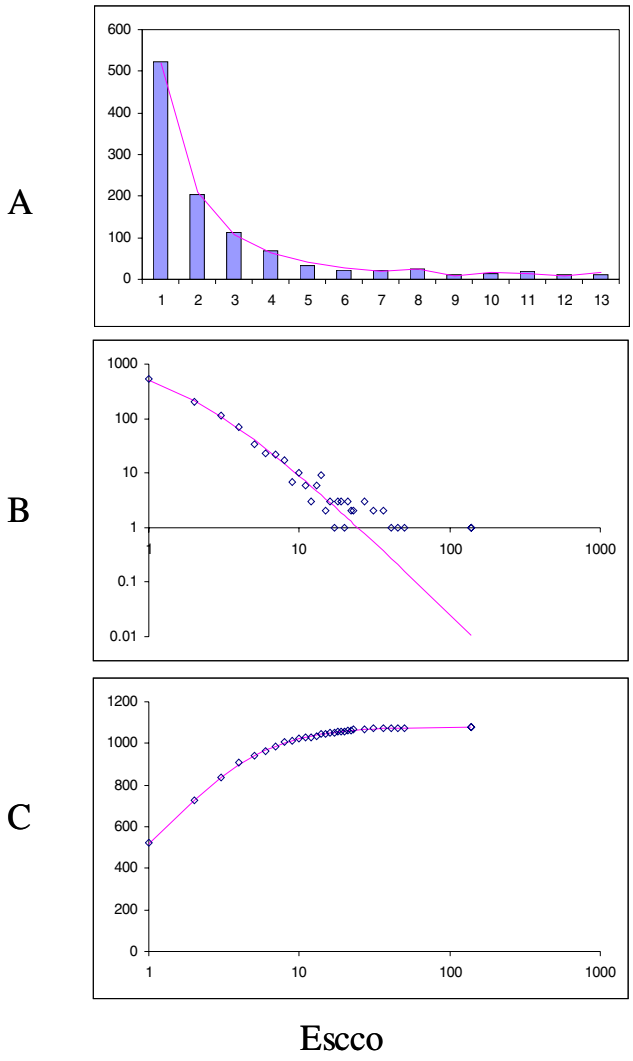

Figure 15
Fit of empirical domain family size distributions to the second-order balanced linear BDIM: the bacterium *Escherichia coli*. The panels and the designations are as in Fig. 6. B. Magenta line: $f_i = 6776\Gamma(i+0.84)/\Gamma(i+3.54)$

We showed that birth-death-innovation models (BDIMs) with different levels of complexity lead to readily distinguishable predictions regarding the distribution of domain family sizes in genomes. In particular, we defined the exact analytic conditions that lead, exactly or asymptotically, to power law distributions, which have recently received ample attention, as they were uncovered in various biological and social contexts [20,25]. In contrast to previous analyses [16,17,30] but in agreement with the results of a recent re-examination [31], we showed that the power law only asymptotically approximates the domain family size distributions.

Three groups of observations made in this work seem to have the greatest potential of enhancing our understanding of genome evolution and, perhaps, evolution of other complex systems. First, we proved that, within the BDIM framework, there is a unique equilibrium state of the system, which is approached exponentially, with respect to time, from any initial state. In this equilibrium state, the number of domain families in each size class remains constant and follows a unique distribution depending on the type and parameters of the BDIM. In particular, power asymptotics emerges when and only when a BDIM is second-order balanced, i.e. the rates of domain birth and death are asymptotically equal. Since we showed that the

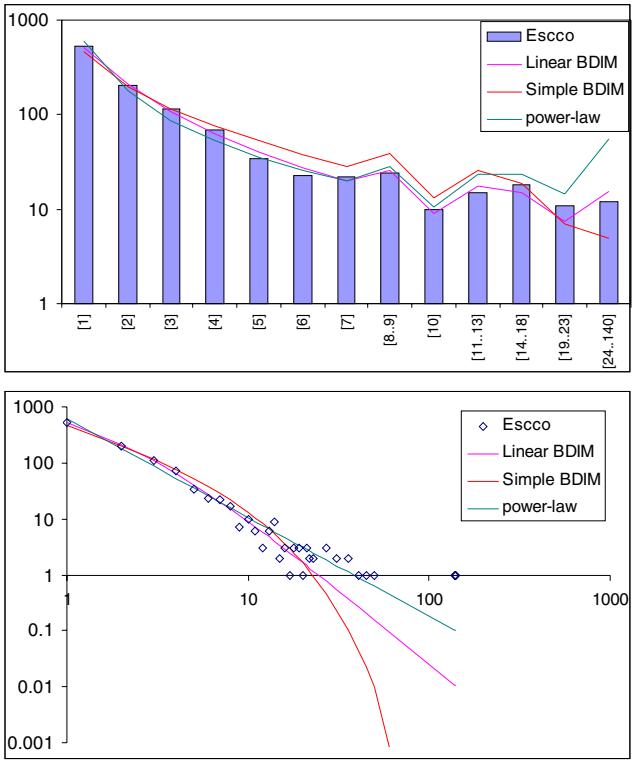

Figure 16
Comparison of different approximations of the empirical domain family size distribution: *Escherichia coli*. Magenta line: second-order balanced linear BDIM, $f_i = 6776\Gamma(i+0.84)/\Gamma(i+3.54)$, Red line: simple BDIM, $f_i = 528 \times 0.87^i/i$, Cyan line: power law, $f_i = 602i^{-1.76}$.

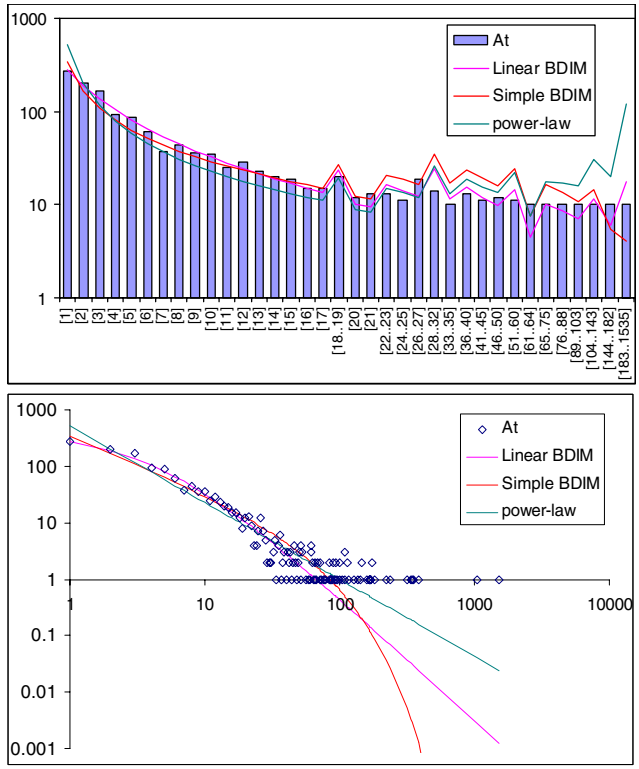

Figure 17
Comparison of different approximations of the empirical domain family size distribution: *Arabidopsis thaliana*. Magenta line: second-order balanced linear BDIM, $f_i = 10750\Gamma(i+3.80)/\Gamma(i+5.98)$, Red line: simple BDIM, $f_i = 344 \times 0.98^i/i$, Cyan line: power law, $f_i = 516i^{-1.36}$.

observed size distributions of domain families in all analyzed genomes indeed tend to power law asymptotics, the results are compatible with the notion that the genomes are close to a steady state with respect to the domain diversity (F_{eq}, the number of distinct domain families at equilibrium, under the using the BDIM convention) and distribution (f_i). Taking a broader biological perspective, this result might indicate that evolving lineages go through lengthy periods of relative stasis when the level of genomic complexity remains more or less the same. Under this view, the stasis epochs are punctuated by relatively short periods of dramatic changes when the complexity either greatly increases (the emergence of eukaryotes is the most obvious case in point) or decreases (e.g. evolution of parasites). These bursts of evolution might be described as transitions between different BDIMs in the parameter space, with some of the trajectories potentially involving non-balanced BDIMs. The analogy between this emerging picture of genome evolution and the punctuated equilibrium concept of species evolution, which has been developed through analysis of the paleontological record [40], is obvious.

Second, we showed that the simplest model that adequately describes the observed domain family size distributions is the second-order balanced linear BDIM; in contrast, simple BDIMs do not show a good fit to the observations. This has potentially important implications for the mode of domain family evolution. Simple BDIMs are based on the notion that the likelihood of duplication (birth) or elimination (death) of a domain is uniform across the genome and does not depend on the size or other characteristics of domain families (the independence assumption). Clearly, under the independence assumption, a duplication (birth) as well as elimination (death) of a domain is more likely to occur in a large family than in a small one, but only because the overall probability of such an event is proportional to the number of family members, whereas the birth (death) rate *per domain* remains the same. The key observation of this work, that the actual domain frequency distributions are well described by a linear but not by a simple BDIM, suggests that the independence assumption is an oversimplification. Instead, the linear BDIM includes parameters that describe the dependence of the per domain birth (death)

rate on the family size. The asymptotics of the theoretical distribution that is the best fit for the actual data is a power law, with the power equal to *a*-*b*-1, where *a* and *b* are the parameters of a linear BDIM. We observed that, for all analyzed genomes, *a*-*b*-1 < -1 (*a* <*b*), which corresponds to "synergy" between domains in a family. Both the domain birth rate and the death rate drop with the increase of the size of a domain family, but the death rate decreases faster (Fig. 5). In general terms, this suggests that small families are more dynamic during evolution than large families. In particular, under the BDIM formalism, innovation contributes only to single-member families (class 1), which have the greatest evolutionary mobility, and either quickly proliferate and are stabilized or perish. An implication of these observations is that, in general, large families are older than small ones. Exceptions to this generalization, i.e. the existence of small, ancient families, probably point to selection for a specific family size; for example, it seems likely that selection acts against proliferation of certain essential proteins, e.g. ribosomal proteins, which typically form single-member families [41]. Another pertinent observation is that the linear BDIM seems to adequately accommodate even the largest of the identified domain families. Lineage-specific expansion of paralogous families appears to be one of the principal modes of organismic adaptation during evolution [13,14,42]. Thus, quantitatively, adaptive family expansion appears to fit within the BDIM framework, although these models do not explicitly incorporate the notion of selection. Of course, for BDIMs, it is irrelevant which families expand, and this choice is determined by selection.

Third, the BDIM equilibrium condition with respect to the total number of domain families, $\nu = \delta_1 f_1$ (ν is the innovation rate, δ_1 is the domain death rate for class 1 families, and f_1 is the number of domain families in class 1) allows us to estimate the ratio between domain innovation rate and the domain death and birth rates. Indeed, according to the above and the definition of a second-order linear BDIM, which is the best fit for the actual data, $\lambda = \delta = \nu/f_1(1+b)$. Since the number of domain families in class 1 (families with only one member) is in the hundreds for each genome, this translates into an innovation rate that is much greater than the duplication or elimination rate *per domain* (Table 1). Such high innovation rates might appear counter-intuitive, but let us note that the duplication rate over all domain families is a number that tends to be nearly identical to ν for small prokaryotic genomes and several-fold greater than ν for large eukaryotic genomes (Table 1). Thus, under the second-order balanced linear BDIM, the likelihood of appearance of a new domain in a genome is close to or several times less than the likelihood of a duplication or elimination of an existing domain. Nevertheless, the finding that the innovation rate is comparable to the overall duplication/elimination rate seems surprising. If the linear BDIM is indeed a realistic evolutionary model, this emphasizes the critical role of innovation in maintaining the balance (steady state) in genome evolution. In prokaryotes, innovation via horizontal gene transfer appears to be particularly extensive [32,39], which might underlie the greater relative innovation rate in these organisms (Table 1).

As already indicated, BDIMs do not explicitly incorporate selection. However, the present analysis shows that only models with precisely balanced domain birth, death and innovation rates can account for the observed distribution of domain family size in each of the analyzed genomes. It seems likely that the balance between these rates is itself a product of selection. There is little doubt that BDIMs will be eventually replaced by more sophisticated formalisms that will more realistically capture the mechanisms of genome evolution. Nevertheless, even the crude modeling described here seems to reveal several potentially interesting and non-trivial aspects of the evolutionary process.

Mathematical Appendix. Proofs of some statements

Proof of Proposition 1

When system (3.1) is solved consecutively from the last equation to the second one, it becomes obvious that the solution is unique up to a constant multiplier.

Next, if $f_i = f_{i-1}\lambda_{i-1}/\delta_i$, $f_{i+1} = f_{i-1}\lambda_{i-1}\lambda_i/(\delta_i\delta_{i+1})$, then the substitution shows that (f_{i-1}, f_i, f_{i+1}) satisfy the i-th equation of system (3.2). Substituting $f_2 = f_1\lambda_1/\delta_2$ in the first equation, we get $f_1 = \nu/\delta_1$ and, consequently, $f_i = \nu\prod_{k=1}^{i-1}\lambda_k / \prod_{k=1}^{i}\delta_k$ for all $i = 2,...N$. By definition, $F_{eq} = \sum_{i=1}^{N} f_i$, so we have (3.3).

Since system (2.2) is linear, the equilibrium state $(f_1,...f_N)$ is asymptotically stable if the real parts of all characteristic values of the matrix

$$\mathbf{A}_N = \begin{pmatrix} -(\lambda_1+\delta_1) & \delta_2 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ \lambda_1 & -(\lambda_2+\delta_2) & \delta_3 & 0 & \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \lambda_{i-1} & -(\lambda_i+\delta_i) & \delta_{i+1} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \dots & 0 & \lambda_{N-2} & -(\lambda_{N-1}+\delta_{N-1}) & \delta_N \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & \lambda_{N-1} & -\delta_N \end{pmatrix}$$

are negative.

The following theorem (see [43]) gives the desired criterion: the real part of all the characteristic values of a real ma-

trix $C = |c_{ij}|$, $i,j = 1,..n$ with non-negative non-diagonal elements are negative if and only if $(-1)^k D_k > 0$ for all $k = 1,..n$, where D_k is the main minor of the matrix C of the k-th order.

To apply this theorem, let us consider the $n \times n$ matrix, $n \le N$

$$B_n = \begin{pmatrix} -(\lambda_1+\delta_1) & \delta_2 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ \lambda_1 & -(\lambda_2+\delta_2) & \delta_3 & 0 & \dots & \dots & \dots & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \lambda_{i-1} & -(\lambda_i+\delta_i) & \delta_{i+1} & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \dots & \dots & \dots & 0 & \lambda_{n-2} & -(\lambda_{n-1}+\delta_{n-1}) & \delta_n \\ 0 & \dots & \dots & \dots & \dots & \dots & 0 & \lambda_{n-1} & -(\lambda_n+\delta_n) \end{pmatrix}$$

It is easy to see that

$$\det \boldsymbol{B}_n = -(\lambda_n + \delta_n)\det \boldsymbol{B}_{n-1} - \lambda_{n-1}\delta_n \det \boldsymbol{B}_{n-2}, \qquad (A1)$$

$$\det \boldsymbol{A}_n = -\delta_n \det \boldsymbol{B}_{n-1} - \lambda_{n-1}\delta_n \det \boldsymbol{B}_{n-2}.$$

Using these equalities, we can prove that for any n

$$\det \boldsymbol{A}_n = (-1)^n \delta_n \delta_{n-1} \dots \delta_2 \delta_1.$$

Indeed,

$$\det \boldsymbol{A}_n = -\delta_n \det \boldsymbol{B}_{n-1} - \lambda_{n-1}\delta_n \det \boldsymbol{B}_{n-2} =$$

$$\delta_n((\lambda_{n-1}+\delta_{n-1}) \det \boldsymbol{B}_{n-2} + \lambda_{n-2}\delta_{n-1} \det \boldsymbol{B}_{n-3}) - \lambda_{n-1}\delta_n \det \boldsymbol{B}_{n-2} =$$

$$\delta_n\delta_{n-1} (\det \boldsymbol{B}_{n-2} + \lambda_{n-2} \det \boldsymbol{B}_{n-3}) = \text{(subsequently using (A1))} =$$

$$(-1)^{n-2}\delta_n\delta_{n-1} \dots \delta_3(\det \boldsymbol{B}_2 + \lambda_2 \det \boldsymbol{B}_1) = (-1)^n \delta_n \delta_{n-1} \dots \delta_2 \delta_1.$$

Further, it is easy to see that for any n

$$\det \boldsymbol{B}_n = \det \boldsymbol{A}_n - \lambda_n \det \boldsymbol{B}_{n-1}.$$

Taking into account that $B_1 = -(\lambda_1 + \delta_1) < 0$ and that the sign of $\det \boldsymbol{A}_n$ coincides with $(-1)^n$, it is easy to prove that

$$\det \boldsymbol{J}_n > \det \boldsymbol{A}_n \text{ if } \det \boldsymbol{A}_n > 0 \text{ and } \det \boldsymbol{J}_n < \det \boldsymbol{A}_n \text{ if } \det \boldsymbol{A}_n < 0.$$

Thus, the sign of $\det \boldsymbol{B}_n$ coincides with the sign of $\det \boldsymbol{A}_n$ and so $(-1)^n \boldsymbol{B}_n > 0$ for all $n = 1,..N$. According to the aforementioned theorem, the real parts of all the characteristic values of a real matrix A_N are negative and so the single equilibrium is asymptotically stable, QED.

Proof of Proposition 2
For simple BDIM (2.1)

$$f_i = \nu \prod_{k=1}^{i-1} \lambda_k / \prod_{k=1}^{i} \delta_k = (\nu/\delta)\theta^{i-1}/i = (\nu/\lambda)\theta^i/i, \text{ so}$$

$$F_{eq} = \sum_{i=1}^{N} f_i = \nu/\lambda \sum_{i=1}^{N} \theta^i/i, \text{ and}$$

$$p_i = f_i/F_{eq} = (\theta^i/i)/\sum_{j=1}^{N} \theta^j/j.$$

If a simple BDIM is balanced, then $\theta = 1$ and so

$$F_{eq} = \nu/\lambda \sum_{j=1}^{N} \theta^j/j.$$

$$p_i = \nu/\lambda\, F_{eq}/i = 1/i\, (\sum_{j=1}^{N} 1/j)^{-1}.$$

Proof of Theorem 1
The condition (3.10) can be rewritten as $\lambda_{i-1}/\delta_i = i^s\theta(1+a/i+O(1/i^2)) = i^s\theta\,(1+a/i)(1+O(1/i^2))$. Thus, we can choose S in such a way that $\prod_{s=S}^{\infty} (1 + O(1/s^2))$ converge, $0 < \prod_{s=S}^{\infty} (1 + O(1/s^2)) < \infty$. It follows that

$$\prod_{s=1}^{j} (\lambda_{s-1}/\delta_s) \sim \Gamma(j)^s\, \theta^j \prod_{s=1}^{j} (1+a/s).$$

According to Proposition 1, $p_i = f_i/F_{eq} \sim \prod_{k=1}^{i-1} \lambda_k / \prod_{k=1}^{i} \delta_k$.

So

$$p_i \sim \prod_{s=2}^{i} (\lambda_{s-1}/\delta_s) \sim \Gamma(i)^s\, \theta^i \prod_{s=2}^{i} (1+a/s) = \Gamma(i)^s\theta^i\, \Gamma(i+a+1)/\Gamma(i+1).$$

Applying the main asymptotic property of Γ-function, i.e. $\Gamma(i+c)/\Gamma(i) \sim i^c$ at large i for any c, we have

$\Gamma(i+a+1)/\Gamma(i+1) \sim i^a$, and so $p_i \sim \Gamma(i)^s\, \theta^i i^a$.

Proofs of Corollaries 1–3
If a discrete random variable ξ has the Pascal distribution, then

$P(\xi = i) \cong 1 / \Gamma(r)\,(1\text{-}q)^r i^{r-1} q^i \sim q^i i^{r-1}$ for large i,

and it becomes evident that, for $a > -1$, equilibrium frequencies p_j of the first-order balanced BDIM follow the Pascal distribution with parameters $(a+1,\theta)$.

If $a = -1$, then $p_i \sim \theta^i/i$ and so p_i follows the truncated logarithmic distribution with parameter θ. If $a = 0$, then $p_j \sim \theta^i$ and p_i follows the geometric distribution.

Further, $p_i \sim i^a$, that is the sequence p_i follows the power distribution with the power a, if and only if $\theta = 1$, that is, if the BDIM is second-order balanced.

Finally, if $\lambda_{i-1}/\delta_i = 1 + O(1/i^2)$, that is, if $\theta = 1$ and $a = 0$, then $p_i \sim$ const; in particular, if $\lambda_{i-1} = \delta_i$ for all i, then, according to Proposition 1, $f_i = \nu$ for all i and $p_i = 1/N$.

Proof of Theorem 2

According to Proposition 1, system (3.1) has the unique solution:

$$f_1 = \nu\delta_1,\ f_i = \nu \prod_{s=1}^{i-1} \lambda_s \Big/ \prod_{s=1}^{i} \delta_s \text{ for all } i = 2,\ldots N.$$ So

$$f_i = \nu/\lambda\theta^i \prod_{s=1}^{i-1} P(s) \Big/ \prod_{s=1}^{i} Q(s),\ i > 1.$$

Applying the Lemma (see below), we get

$f_i \cong C\nu/\lambda\ \theta^i\,\Gamma(i)^\eta i^{\rho-\beta}$, as $i\to\infty$,

where the constant $C = \prod_{k=1}^{m} [(\Gamma(1+b_k))^\beta_k] \Big/ \prod_{k=1}^{n} [\Gamma(1+a_k)^\alpha_k]$.

Lemma. Let $P(j) = \prod_{k=1}^{n} (j+a_k)^\alpha_k$, $Q(j) = \prod_{k=1}^{m} (j+b_k)^\beta_k$, where a_k, b_k are positive. Let us denote

$$\eta = \sum_{k=1}^{n} \alpha_k - \sum_{k=1}^{m} \beta_k,\ \rho = \sum_{k=1}^{n} a_k\,\alpha_k - \sum_{k=1}^{m} b_k\beta_k,\ \beta = \sum_{k=1}^{m} \beta_k.$$

Then with fixed j

$$N(j) = \prod_{s=1}^{j-1} P(s) \Big/ \prod_{s=1}^{j} Q(s) \cong C\,\Gamma(j)^\eta j^{(\rho-\beta)}$$

as $j\to\infty$, where

$$C = \prod_{k=1}^{m} [(\Gamma(1+b_k))^\beta_k] \Big/ \prod_{k=1}^{n} [\Gamma(1+a_k)^\alpha_k].$$

Proof.

$$\prod_{s=1}^{j-1} (s+a_k)^\alpha_k = [\Gamma(j+a_k) / \Gamma(1+a_k)]^\alpha_k,$$

$$\prod_{s=1}^{j} (s+b_k)^\beta_k = [\Gamma(j+1+b_k) / \Gamma(1+b_k)]^\beta_k,$$ so

$$N(j) = \{\prod_{k=1}^{n} [\Gamma(j+a_k) / \Gamma(1+a_k)]^\alpha_k\} / \{\prod_{k=1}^{m} [\Gamma(j+1+b_k)/ \Gamma(1+b_k)]^\beta_k\} =$$

$$C \prod_{k=1}^{n} [(\Gamma(j+a_k))^\alpha_k] / \prod_{k=1}^{m} [(\Gamma(j+1+b_k))^\beta_k]$$

where

$$C = \prod_{k=1}^{m} [(\Gamma(1+b_k))^\beta_k] / \prod_{k=1}^{n} [\Gamma(1+a_k)^\alpha_k].$$

Let us use the known asymptotic relation

$\Gamma(t+a)/\Gamma(t) \cong t^a$ with $t\to\infty$.

Thus we have

$$\prod_{k=1}^{n} [(\Gamma(j+a_k))^\alpha_k] / \prod_{k=1}^{m} [(\Gamma(j+1+b_k))^\beta_k]$$

$$(\Gamma(j))^\eta \prod_{k=1}^{n} [(\Gamma(j+a_k) / \Gamma(j))^\alpha_k] / \prod_{k=1}^{m} [(\Gamma(j+1+b_k) / \Gamma(j))^\beta_k] \cong$$

$$(\Gamma(j))^\eta j^{[\prod_{k=1}^{n} a_k\,\alpha_k]} / j^{[\prod_{k=1}^{m} (b_k+1)\beta_k]} =$$

$$(\Gamma(j))^\eta j^{(\rho-\beta)},$$

and Lemma is proved.

Proof of Proposition 3

It follows from the proof of the Lemma that

$$f_i = C\nu/\lambda\ \theta^i \prod_{k=1}^{n} [(\Gamma(j+a_k))^{\alpha_k}] / \prod_{k=1}^{m} [(\Gamma(j+1+b_k))^{\beta_k}]$$

for $i > 1$,

where $C = \prod_{k=1}^{m} [(\Gamma(1+b_k))^{\beta_k}] / \prod_{k=1}^{n} [(\Gamma(1+a_k))^{\alpha_k}]$.

Let us show that f_1 can be expressed by the same formula if $i = 1$. Indeed,

$$C\nu/\delta \prod_{k=1}^{n} [(\Gamma(1+a_k))^{\alpha_k}] / \prod_{k=1}^{m} [(\Gamma(1+1+b_k))^{\beta_k} =$$

$$\nu/\delta \left(\prod_{k=1}^{m} (\Gamma(1+b_k))^{\beta_k} / \prod_{k=1}^{n} (\Gamma(1+a_k))^{\alpha_k}\right) \left(\prod_{k=1}^{n} (\Gamma(1+a_k))^{\alpha_k} / \prod_{k=1}^{m} (\Gamma(2+b_k))^{\beta_k} =\right.$$

$$\nu/\delta \left(\prod_{k=1}^{m} (\Gamma(1+b_k))^{\beta_k} / \prod_{k=1}^{m} (\Gamma(2+b_k))^{\beta_k} = \nu/\delta \left(\prod_{k=1}^{m} (1+b_k))^{\beta_k} = f_1\right.\right.$$

Thus,

$$F_{eq} = C\nu/\delta \left(\sum_{j=1}^{N} \theta^{j-1} \prod_{k=1}^{n} (\Gamma(j+a_k))^{\alpha_k} / \prod_{k=1}^{m} (\Gamma(j+1+b_k))^{\beta_k}\right).$$

QED.

Contributions of individual authors

GPK developed most of the mathematical formalism and wrote the draft of the mathematical part of the manuscript; YIW performed the identification of domain in sequenced genomes and the statistical analysis of the resulting distributions and wrote the draft of the corresponding part of the manuscript; FSB proved some of the theorems; AYR largely incepted the work and contributed to the formulation of the models; EVK contributed to the inception of the work and the formulation of the models, gave the biological interpretation of the results, wrote the background and discussion sections and extensively edited the entire manuscript.

Acknowledgements

We thank Alexei Kondrashov, Alexei Ogurtzov, and Vladimir Ponomarev for critical reading of the manuscript and the Koonin group members for helpful discussions.

References

1. Koonin EV, Aravind L, Kondrashov AS: **The impact of comparative genomics on our understanding of evolution.** *Cell* 2000, **101:**573-576
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, *et al*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409:**860-921
3. Dacks JB, Doolittle WF: **Reconstructing/deconstructing the earliest eukaryotes: how comparative genomics can help.** *Cell* 2001, **107:**419-425
4. Chervitz SA, Aravind L, Sherlock G, Ball CA, Koonin EV, Dwight SS, Harris MA, Dolinski K, Mohr S, Smith T, Weng S, Cherry JM, Botstein D: **Comparison of the complete protein sets of worm and yeast: orthology and divergence.** *Science* 1998, **282:**2022-2028
5. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W, Cherry JM, Henikoff S, Skupski MP, Misra S, Ashburner M, Birney E, Boguski MS, Brody T, Brokstein P, Celniker SE, Chervitz SA, Coates D, Cravchik A, Gabrielian A, Galle RF, Gelbart WM, George RA, Goldstein LS, Gong F, Guan P, Harris NL, Hay BA, Hoskins RA, Li J, Li Z, Hynes RO, Jones SJ, Kuehl PM, Lemaitre B, Littleton JT, Morrison DK, Mungall C, O'Farrell PH, Pickeral OK, Shue C, Vosshall LB, Zhang J, Zhao Q, Zheng XH, Zhong F, Zhong W, Gibbs R, Venter JC, Adams MD, Lewis S: **Comparative genomics of the eukaryotes.** *Science* 2000, **287:**2204-2215
6. Koonin EV, Tatusov RL, Rudd KE: **Sequence similarity analysis of *Escherichia coli* proteins: functional and evolutionary implications.** *Proc Natl Acad Sci U S A* 1995, **92:**11921-11925
7. Brenner SE, Hubbard T, Murzin A, Chothia C: **Gene duplications in *H. influenzae*.** *Nature* 1995, **378:**140
8. Labedan B, Riley M: **Widespread protein sequence similarities: origins of Escherichia coli genes.** *J Bacteriol* 1995, **177:**1585-1588
9. Tatusov RL, Mushegian AR, Bork P, Brown NP, Hayes WS, Borodovsky M, Rudd KE, Koonin EV: **Metabolism and evolution of Haemophilus influenzae deduced from a whole-genome comparison with *Escherichia coli*.** *Curr Biol* 1996, **6:**279-291
10. Ponting CP, Schultz J, Milpetz F, Bork P: **SMART: identification and annotation of domains from signalling and extracellular protein sequences.** *Nucleic Acids Res* 1999, **27:**229-232
11. Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2002, **30:**276-280
12. Marchler-Bauer A, Panchenko AR, Shoemaker BA, Thiessen PA, Geer LY, Bryant SH: **CDD: a database of conserved domain alignments with links to domain three-dimensional structure.** *Nucleic Acids Res* 2002, **30:**281-283
13. Jordan IK, Makarova KS, Spouge JL, Wolf YI, Koonin EV: **Lineage-specific gene expansions in bacterial and archaeal genomes.** *Genome Res* 2001, **11:**555-565

14. Lespinet O, Wolf YI, Koonin EV, Aravind L: **The role of lineage-specific gene family expansion in the evolution of eukaryotes.** *Genome Res* 2002, **12:**1048-1059
15. Aravind L, Watanabe H, Lipman DJ, Koonin EV: **Lineage-specific loss and divergence of functionally linked genes in eukaryotes.** *Proc Natl Acad Sci U S A* 2000, **97:**11319-11324
16. Huynen MA, van Nimwegen E: **The frequency distribution of gene family sizes in complete genomes.** *Mol Biol Evol* 1998, **15:**583-589
17. Qian J, Luscombe NM, Gerstein M: **Protein family and fold occurrence in genomes: power-law behaviour and evolutionary model.** *J Mol Biol* 2001, **313:**673-681
18. Rzhetsky A, Gomez SM: **Birth of scale-free molecular networks and the number of distinct DNA and protein domains per genome.** *Bioinformatics* 2001, **17:**988-996
19. Wuchty S: **Scale-free behavior in protein domain networks.** *Mol Biol Evol* 2001, **18:**1694-1702
20. Barabasi AL: *Linked: The New Science of Networks. New York: Perseus Pr* 2002
21. Barabasi AL, Albert R: **Emergence of scaling in random networks.** *Science* 1999, **286:**509-512
22. Albert R, Barabasi AL: **Statistical mechanics of complex networks.** *Reviews of Modern Physics* 2002, **74:**47-97
23. Albert R, Jeong H, Barabasi AL: **Error and attack tolerance of complex networks.** *Nature* 2000, **406:**378-382
24. Amaral LA, Scala A, Barthelemy M, Stanley HE: **Classes of small-world networks.** *Proc Natl Acad Sci U S A* 2000, **97:**11149-11152
25. Gisiger T: **Scale invariance in biology: coincidence or footprint of a universal mechanism?** *Biol Rev Camb Philos Soc* 2001, **76:**161-209
26. Jeong H, Mason SP, Barabasi AL, Oltvai ZN: **Lethality and centrality in protein networks.** *Nature* 2001, **411:**41-42
27. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabasi AL: **The large-scale organization of metabolic networks.** *Nature* 2000, **407:**651-654
28. Dorogovtsev SN, Mendes JF: **Scaling properties of scale-free evolving networks: Continuous approach.** *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 2001, **63:**056125
29. Krapivsky PL, Redner S: **Organization of growing random networks.** *Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Topics* 2001, **63:**066123
30. Luscombe N, Qian J, Zhang Z, Johnson T, Gerstein M: **The dominance of the population by a selected few: power-law behaviour applies to a wide variety of genomic properties.** *Genome Biol.* 2002, **3:**research0040.0041-0040.0047
31. Kuznetsov VA: **Statistics of the numbers of transcripts and protein sequences encoded in the genome.** *In: 'Computational and Statistical Approaches to Genomics (Edited by: Zhang W, Shmulevich I) Boston: Kluwer* 2002, 125-171
32. Koonin EV, Makarova KS, Aravind L: **Horizontal gene transfer in prokaryotes: quantification and classification.** *Annu Rev Microbiol* 2001, **55:**709-742
33. Feller W: *An introduction to probability theory and its application. New York: Wiley* 1967–1968
34. Ijuri Y, Simon HA: *Skew distributions and the sizes of business firms. Amsterdam, New York, Oxford: North-Holland Publishing Company* 1977
35. Gihman II, Skorohod AV: *The theory of stochastic processes. New-York, Heidelberg, Berlin: Springer-Verlag* 1975
36. Johnson NL, Kotz S, Kemp AW: *Univariate discrete distributions. New York: Wiley* 1992
37. Henrici P: *Applied and computational complex analysis. New York: Wiley* 1986
38. Wolf YI, Grishin NV, Koonin EV: **Estimating the number of protein folds and families from complete genome data.** *J Mol Biol* 2000, **299:**897-905
39. Lawrence JG, Ochman H: **Amelioration of bacterial genomes: rates of change and exchange.** *J Mol Evol* 1997, **44:**383-397
40. Gould SJ: *The Structure of Evolutionary Theory. Cambridge, MA: Harvard Univ. Press* 2002
41. Tatusov RL, Galperin MY, Natale DA, Koonin EV: **The COG database: a tool for genome-scale analysis of protein functions and evolution.** *Nucleic Acids Res* 2000, **28:**33-36
42. Remm M, Storm CE, Sonnhammer EL: **Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.** *J Mol Biol* 2001, **314:**1041-1052
43. Gantmacher FR: *The theory of matrices. New York: Chelsea Publishing Company* 1989